High-dimensional data: introduction

Wessel van Wieringen
w.n.van.wieringen@vu.nl

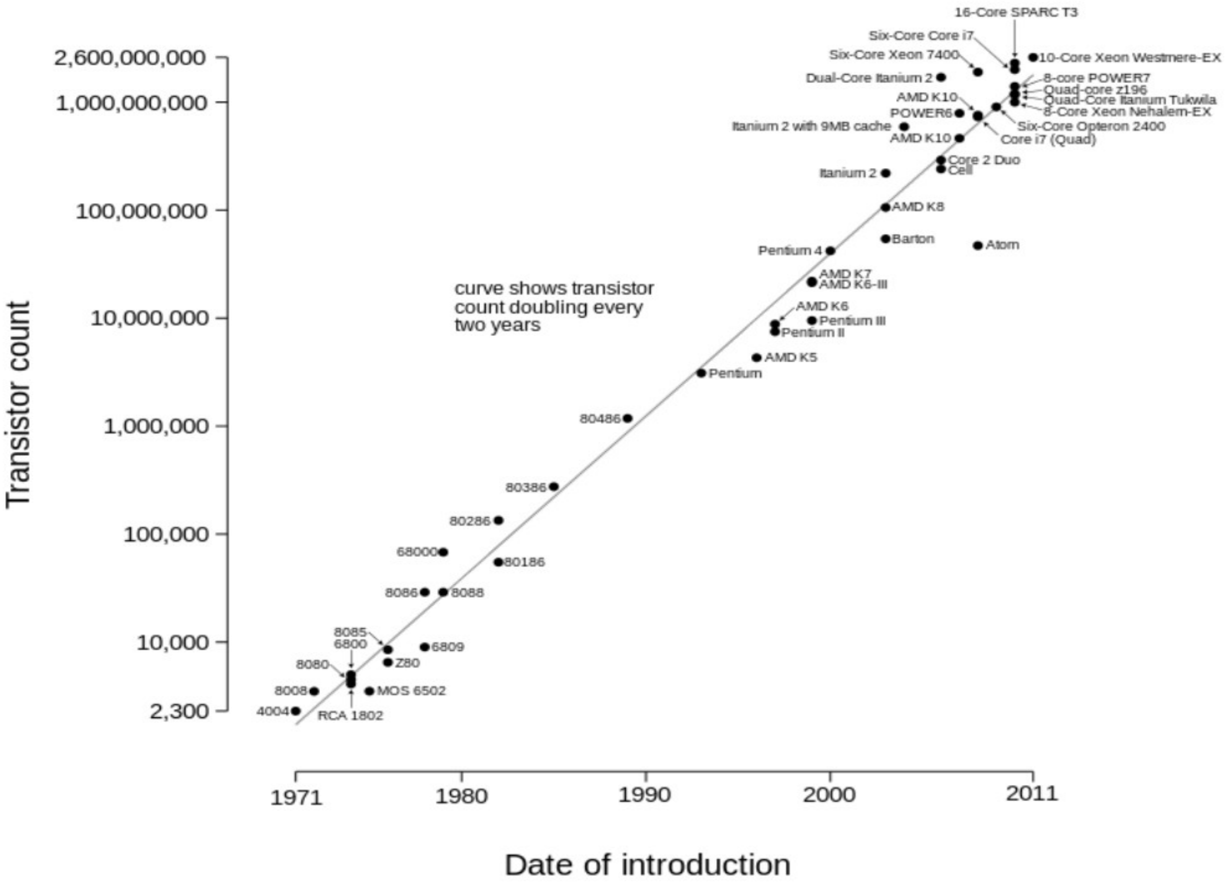
Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

How did we end up here?

Moore's law
The number of transistors in a dense integrated circuit doubles approximately every two years.

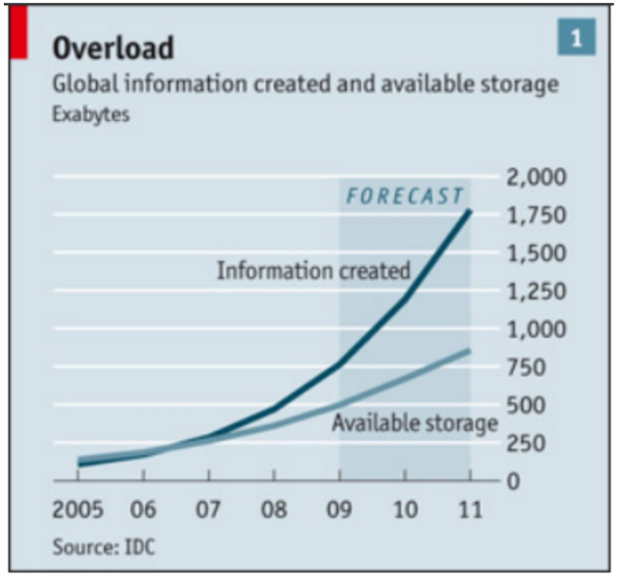


source: wikipedia

How did we end up here?

Data deluge

“... the quantity of information in the world is soaring. According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still.”

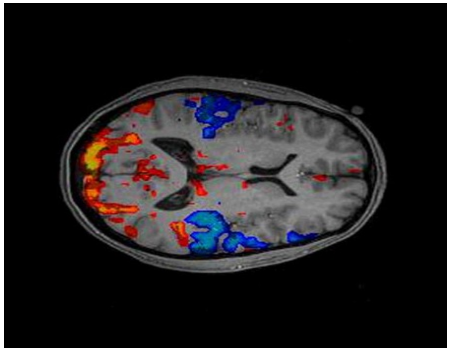


source: The economist

How did we end up here?

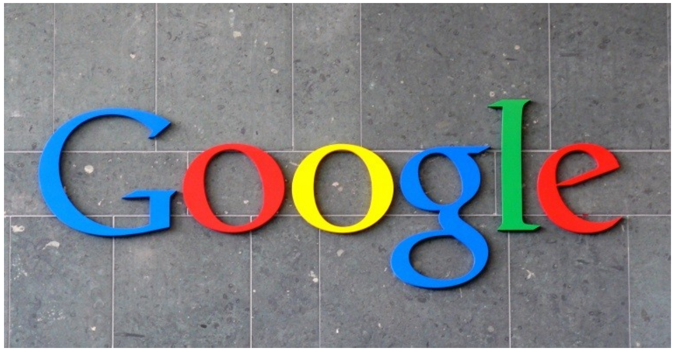
Examples

→ Brain image data (fMRI / EEG)

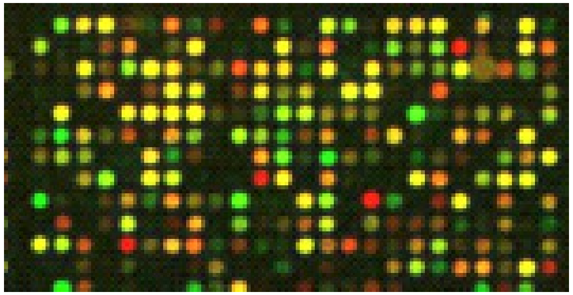


→ Movie database

→ Google search data



→ Microarrays

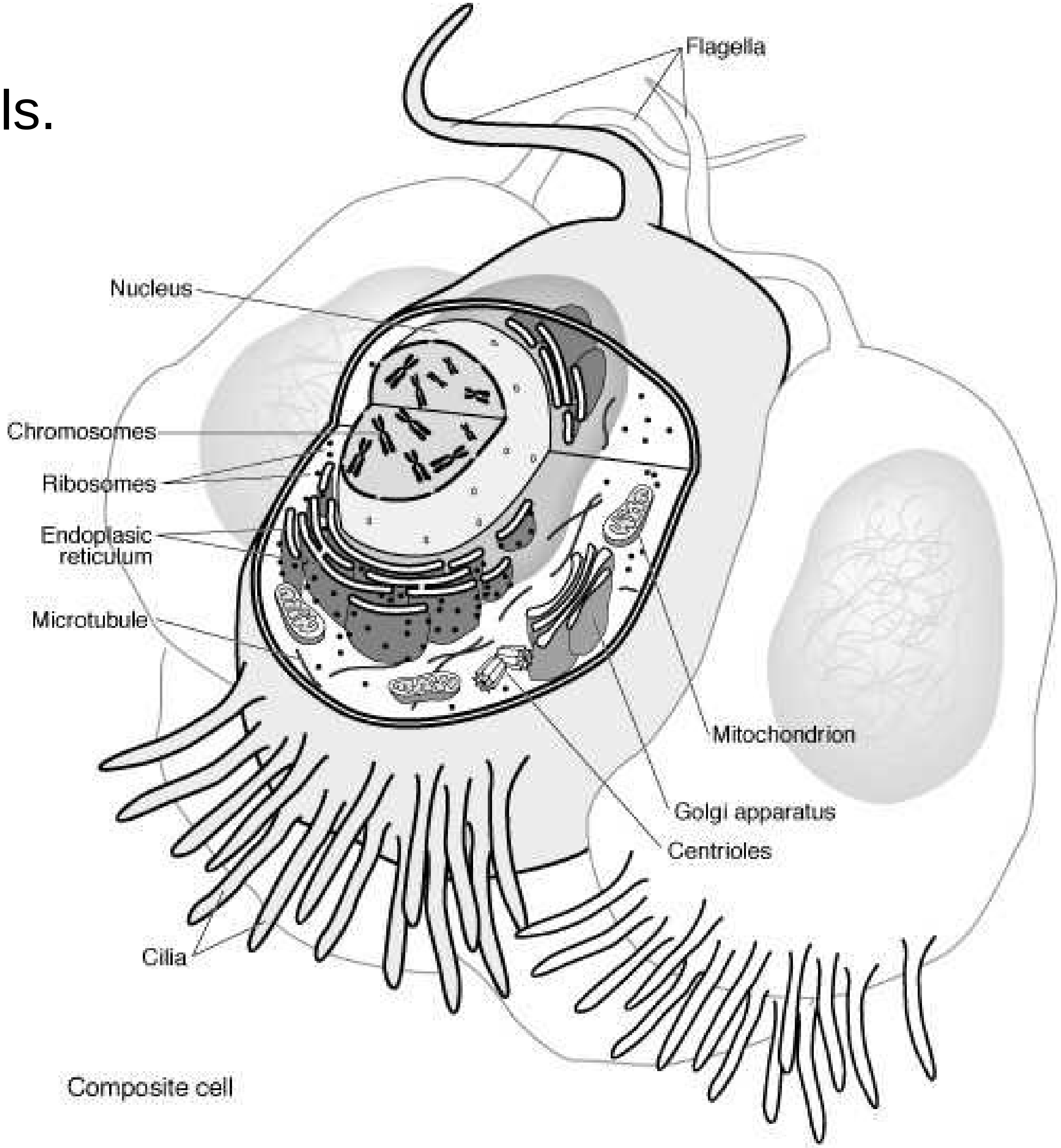


A minimum of biology

Organisms are made of cells.

A *cell* is the smallest possible independent living unit. The cell contains a complete copy of the organisms genome.

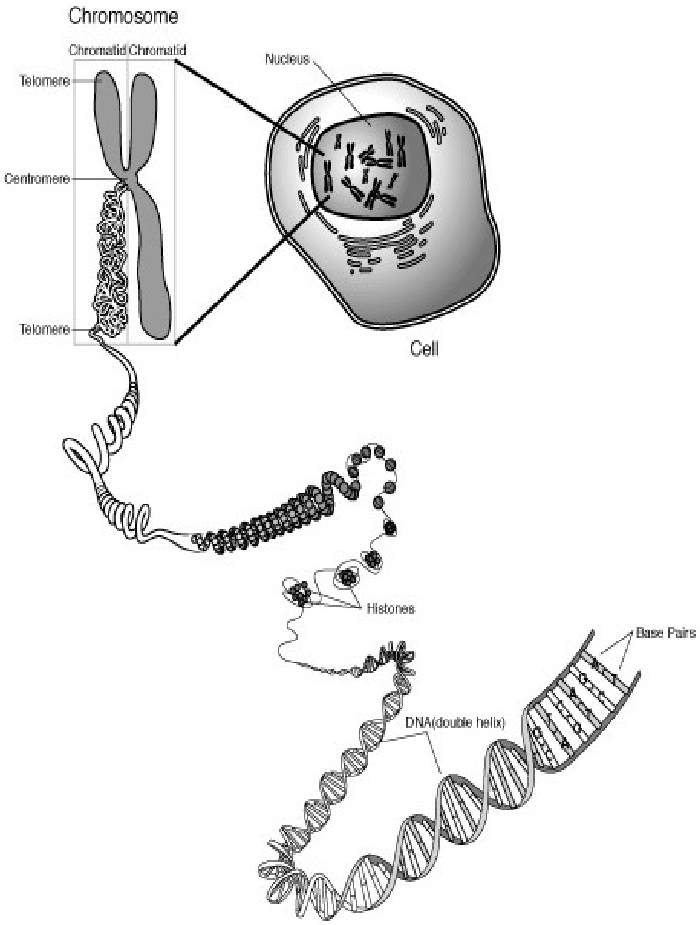
The *genome* is the total genetic constitution of an organism, the full haploid set of chromosomes with all its genes.



Composite cell

A minimum of biology

A *chromosome* is one of a set of threadlike molecular structures composed of compressed *DNA*, that carry the genes which determine an individual’s heriditary traits.

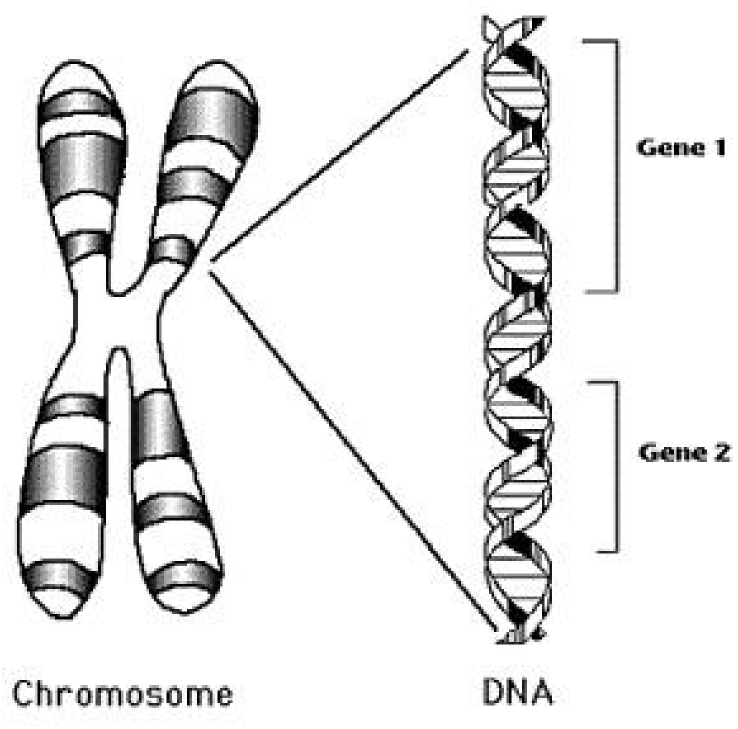


Conceptually, *DNA* is an information carrier, information necessary for the functioning of cells and encoded in molecular units called genes.

On the molecular level *DNA* is a double-stranded polymer composed of four basic molecular units called nucleotides.

A minimum of biology

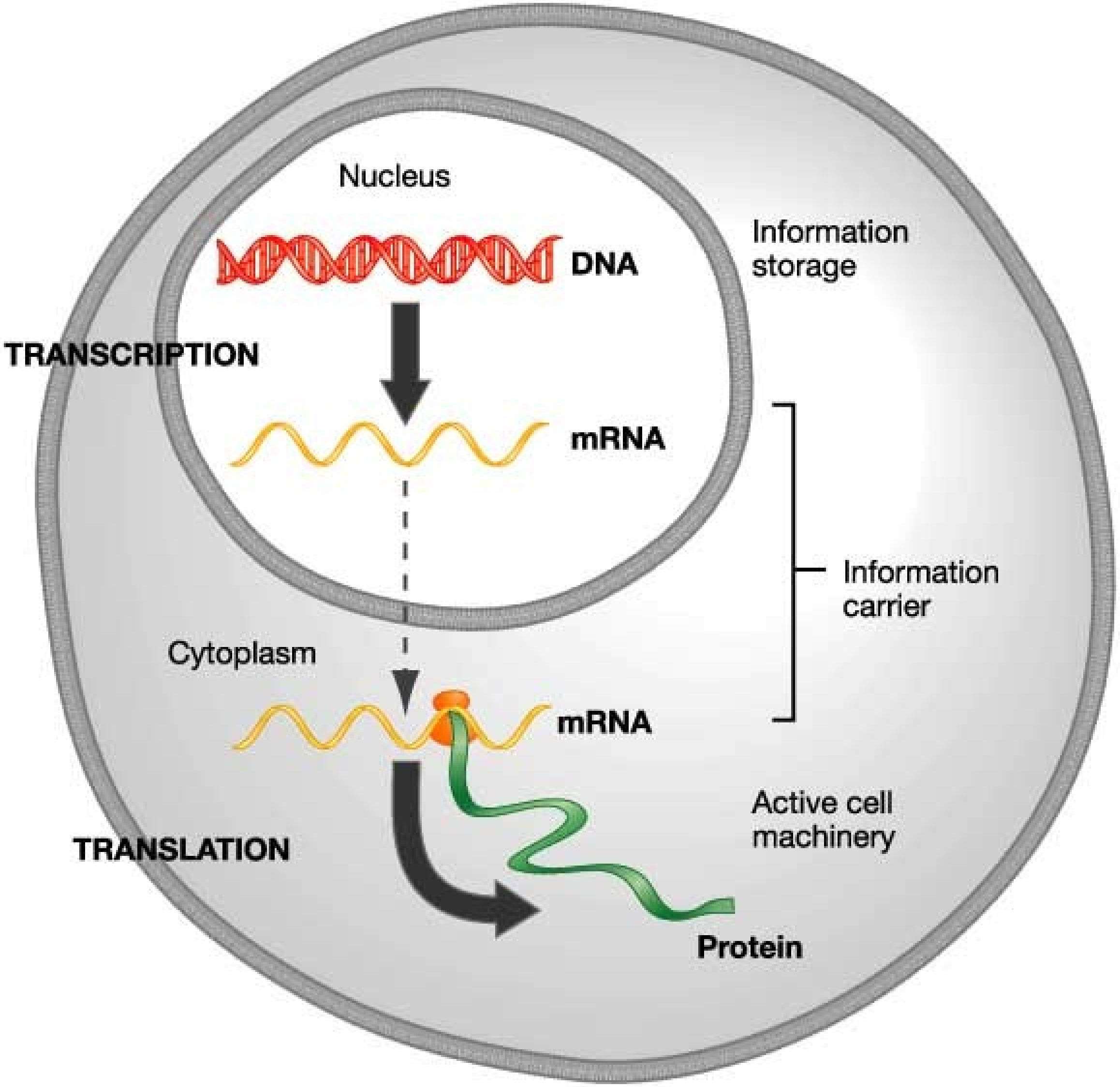
A *gene* is the basic physical unit of heredity: a linear sequence of nucleotides, as a segment of DNA located on a chromosome, that provides the coded instruction for one polypeptide chain.



A minimum of biology

Central Dogma of Molecular Biology describes the information transfer process that leads from the information encoded in DNA to the proteins in the cell.

Three steps are discerned:
1) Replication
2) Transcription
3) Translation

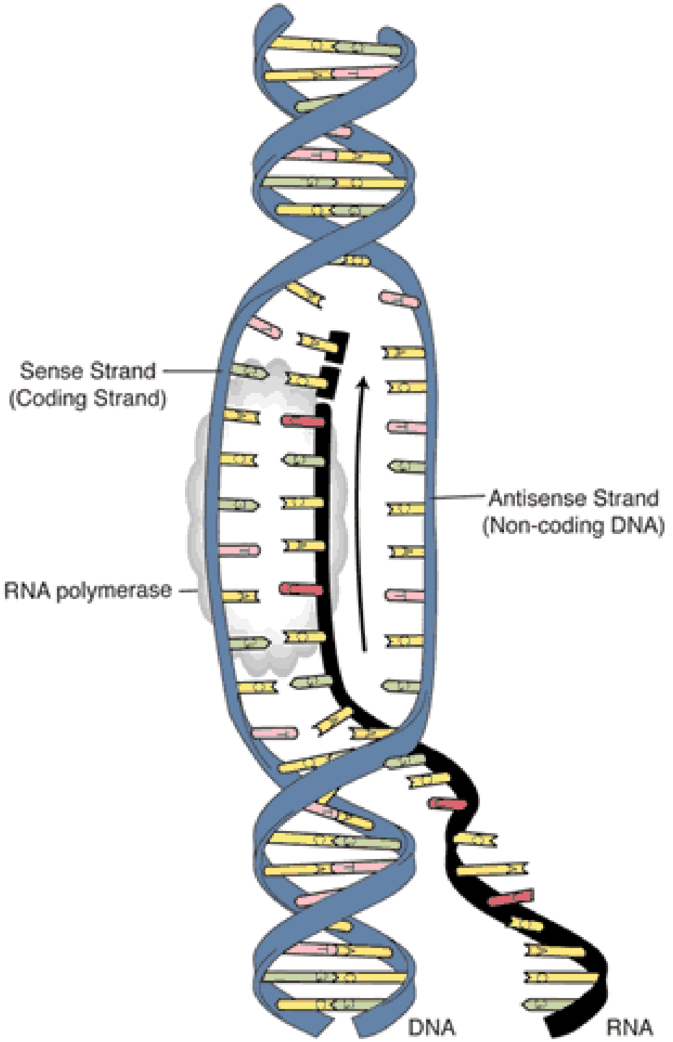


A minimum of biology

Transcription is the synthesis of *mRNA* (nucleic acid like DNA, but single-stranded) from the DNA in the nucleus.

The mRNA is transported to the cytoplasm and used to synthesize protein.

Jargon
A gene is said to be *expressed* if the product it encodes for has been formed.



A minimum of biology

Molecular biology aims to understand the molecular processes that occur in the cell. That is, which molecules present in the cell interact, and how is this coordinated?

For many cellular process, it is unknown which genes play what role.

Solution

Simply measure (the expression of) all genes ...
… and later sort out which are relevant.

Microarrays

Microarray

- Conceptually: a measurement device.

Gene expression arrays measure the expression of genes (which genes are expressed and to what extent).

In fact, it measures mRNA which is related – through the transcription process – to the expression of genes.

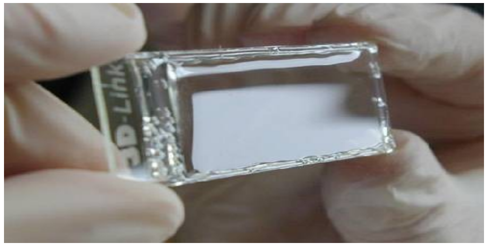
Other types of microarrays measure:

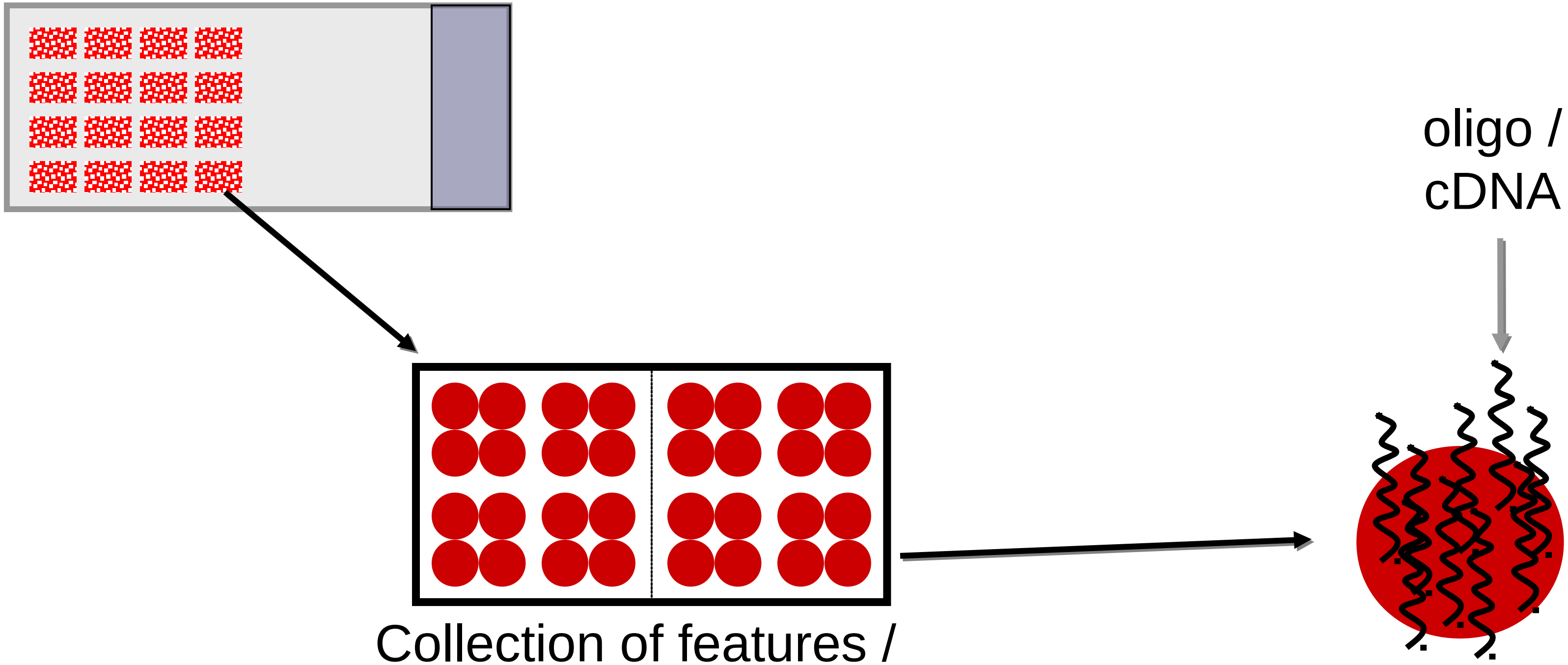
- SNPs
- DNA copy number
- methylation
- ...

Microarrays

Microarray

- Physically: a glass slide





Microarray

oligo / cDNA

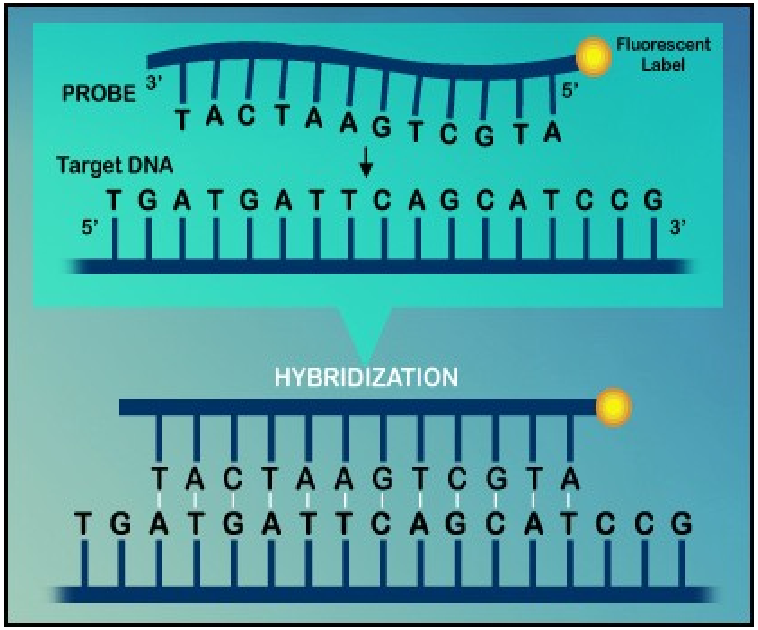
Collection of features / probes

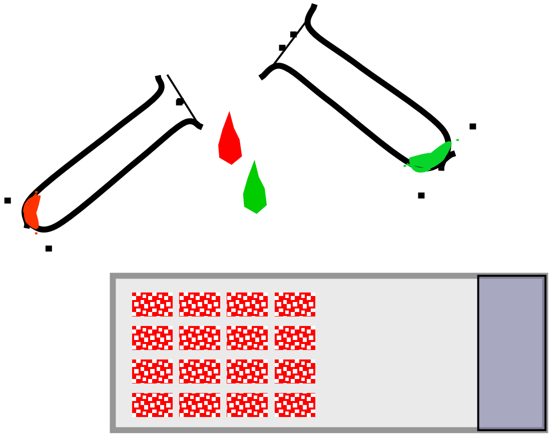
Feature / probe

Microarrays

Hybridization

The preferential binding of gene sequences to complementary sequences.

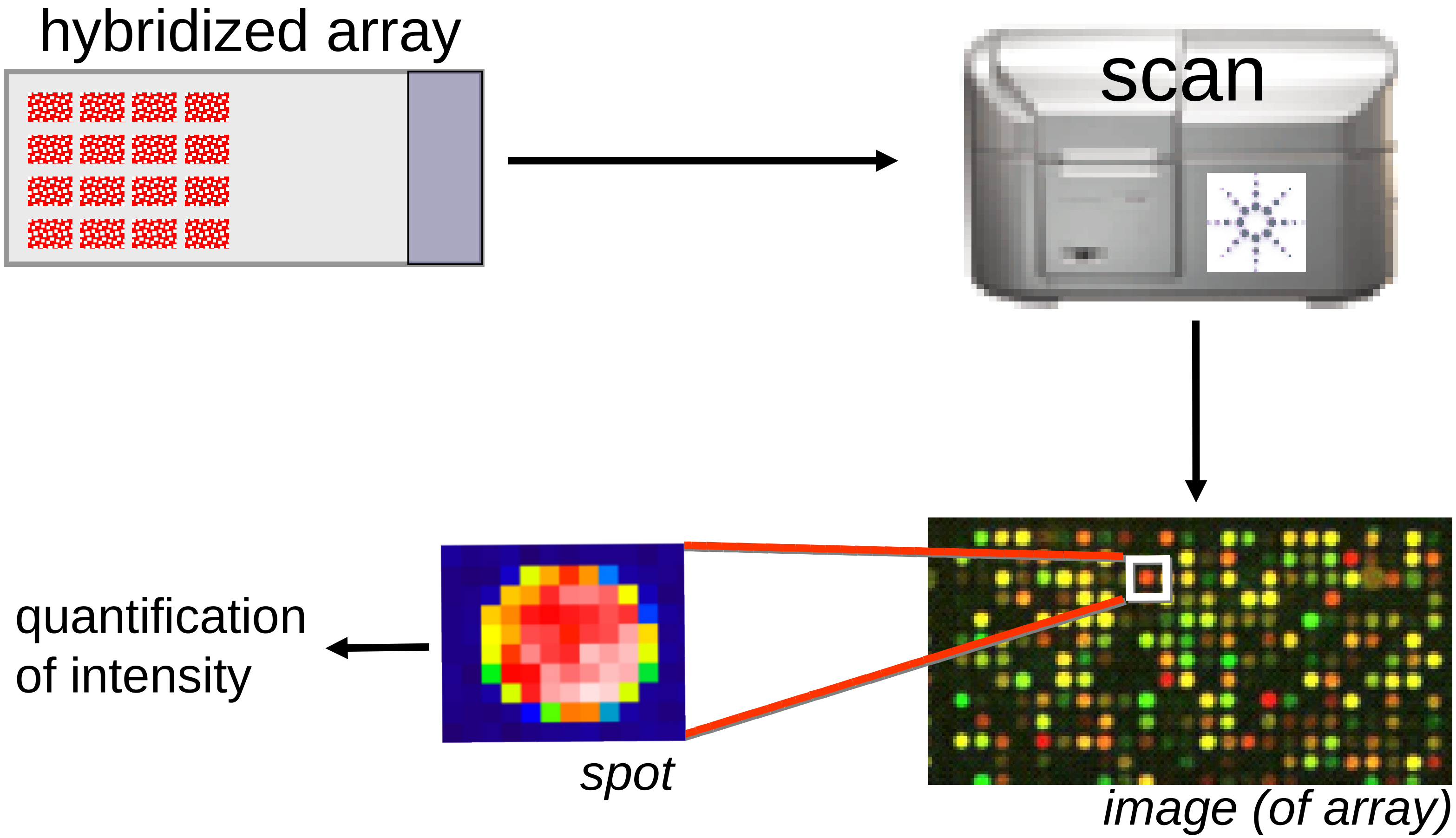




Picture: http://members.cox.net/amgough/

Microarrays

An image of the microarray is generated



hybridized array

scan

image (of array)

spot

quantification of intensity

Microarrays

Output per hybridization:

- File of 44Mb
- ~44000 rows
- ~100 columns
- Annotation information
- Quality metrics
- Biological signal (in various forms)
- Background signal (in various forms)

```
[10,]   1.3706866  -3.1815053   1.0945986 -2.2518298  -0.3100496  -0.6686597  -0.6540535  -1.9287335 -3.2379697  -1.2578805  -3.7475174
[11,]   1.0482100   1.3511300   1.2707900  0.9512500   1.3523800   1.3059000   1.1234800   0.8849800  1.2081000   1.5833800   1.1002200
[12,]  -0.2838600  -0.2799200  -0.0740500 -0.0607000  -0.1874500  -0.0489800  -0.9407800   0.2707600  0.0189300  -0.1034100   0.5968100
[13,]  -0.0826900  -0.4920800   0.1461000  1.0789000  -1.4266800   0.6920100  -0.3478800   1.0291000 -0.7605600  -1.2158300   0.0005400
[14,]   1.2655600   0.6419300   1.4654800  1.8207700   2.0073600   1.4918500   1.4782200  -0.2940800  1.9542600   1.2704300   1.2622500
[15,]   1.0793800   0.8870500   1.2652700  1.3917500   1.5701100   1.3536700   1.3709700   1.0549900  1.1460900   1.3019800   1.0526800
[16,]  -0.5766680  -1.7953233   0.4123687 -1.3000260  -0.3914956  -0.7176529  -1.9670290  -1.2006734 -0.5781646  -0.1262439  -1.9169129
[17,]   0.7748900   0.4785400   0.6538900  0.5344200   0.6830700   0.6679600   0.6442800   0.7885300  1.0363900   0.7424500   0.9729600
[18,]   0.2745500  -0.0585100   0.3185400  0.2388900  -0.3607500  -0.0649300   0.1958400   0.4509100 -0.1332400   0.2465000   0.0005400
[19,]  -1.7879147  -1.0191236  -1.5552526 -3.5924412  -1.7454766  -0.7730166  -0.3718211  -0.5105483 -1.2629439  -2.9315562  -0.6414949
[20,]   0.5260500   0.3957500   0.6546800  0.5320500   0.2394300   0.6618500   0.5165000   0.6978100  0.2828800   0.3952900   0.3588100
[21,]  -1.4576900  -0.0863700  -1.4622700 -1.4071500  -1.4266800  -1.2171900  -1.3738600  -1.3683200 -1.1429700   0.6344900  -0.8615000
[22,]  -9.6719893   2.2601265   7.7628864 -3.4866526   4.3207356  -7.2708633 -13.5159651   2.1101495  1.2602673  -6.0691419 -12.5633527
[23,]  -1.6251345  -1.7400089  -0.8286638 -1.8216632  -1.4516043  -1.0365167  -0.4819336  -0.8287533 -1.4291445  -2.1557932  -0.3753139
[24,]   0.2773600  -0.8494900  -0.1880900  0.7294600   0.8723400   0.2823900   0.5251000  -0.1515900  0.4077000  -1.2158300  -0.2254300
[25,]   0.9831800   1.3916500   1.4639100  0.8695400   0.9272300   1.4933500   1.7161000   1.3445300  1.3880900   1.4714300   1.1528100
[26,]  -0.4879006   2.3522517   3.9431369 -0.7394428  -4.5868684  -4.9465152  -4.8069718  -3.3983413 -4.1992253   0.4114557  -2.9567431
[27,] -12.3719528   1.2538366  -6.3306388 -8.3371014   0.8558988   0.7665416  -2.1466177  -7.3020149 -1.8094012   2.9542900   0.9333962
[28,]  -1.0775500  -1.3942000  -0.8933400 -0.9148300  -1.4266800  -0.5972500  -0.7218800   0.1395300 -0.8057400  -1.2158300  -1.2813700
[29,]  -5.4636983  -8.7262323 -11.6165240 -9.5747269  -2.9136981  -5.7593651  -5.6418661  -5.9271453 -7.5683810  -7.1324658  -7.1897210
[30,]   0.8047400   1.1665100   0.9457400  0.7891100   0.2988400   1.2077600   0.9021200   0.9256000  0.7943100   1.1191800   1.1347800
[31,]   1.6280900   1.3590700   1.7875200  1.6325900   1.5195200   1.8001000   1.7636900   1.9085500  1.5812000   1.4344900   1.6826600
[32,] -15.9177707  -4.4406363   4.7738056 -7.3396969  -3.9618396 -10.7848498 -10.1883121 -10.6706668  3.8301984 -15.6259941 -12.0914823
[33,]   0.6881200   1.2516600   0.7484200  0.9140500   1.4083300   0.8687400   0.6728900   0.4837400  1.0766700   1.1625200   1.1341500
[34,]   0.3385600   0.5253400  -0.0596600 -0.8249000   0.8082000   0.9212000  -0.7312200  -0.7411500  0.8827000   0.8231600   1.1157400
[35,]  -0.7821400   0.0807700  -0.3072600 -0.7021500   0.0662200  -0.4280100  -0.2200700   0.0004800  0.3340100  -0.7439600   0.4983800
[36,]  -1.4576900  -1.0118900   0.1766100 -0.6631400  -1.4266800  -0.8595600  -0.0839000  -0.9927200 -0.5058000   0.0757300  -0.4963000
[37,]   0.2327100   1.0574200   0.7632300  0.8612200   1.4853800   0.7473100   0.3884400   0.1653000  1.1433600   1.0985900   1.1872300
[38,]  -0.1015800  -0.4637700   0.1461000 -0.0856000  -0.7132200  -0.2647300   0.2773900  -0.2664600 -0.0805300  -0.2425100  -0.2511800
[39,]  -0.0480300   0.4828000  -0.3608500 -0.2984100   0.6687000  -0.2256400  -0.0538900  -0.5666200  0.4772200   0.1803300   0.0307900
[40,]   0.6457300  -0.3533100   0.4244100 -0.1821700   0.3117000   0.1769600   0.3567800  -0.1983400  0.0130500  -1.2158300   0.4621500
[41,] -13.2327629  -7.4784300  -2.3736338  3.6926116  -0.1132352  -9.2827358 -10.1963165 -10.2116792 -4.4631231 -18.6168573  -6.8622131
[42,]  -1.1448447   1.6734572   0.2015480 -0.2543426  -3.2946844  -2.1751983  -0.1855370  -0.6094323 -1.8925225  -0.2569482  -0.6987012
[43,]  -1.0331500   7.8505726  -3.4200628 -0.5226154 -11.0591912 -13.9737873  -7.0726104 -18.2617373  5.1637078   7.8912059   1.2117123
[44,]  -5.2197951 -11.8997017  -8.4754373 -6.5459542 -11.5095411 -11.7889469  -5.3662040  -4.9429571 -6.7202350   1.1376922  -8.4510359
[45,]  -0.9708282  -6.1642629   0.8067003 -1.4437932  -1.6959041  -0.1198638  -1.4731559  -1.3784706 -4.1129785  -0.1576396  -2.8251738
[46,]  -2.1051862   0.6287777  -3.4271863  0.7024729  -1.8750100  -0.6556041   8.5340851  -2.6672457 -1.5931683  -2.3232690  -2.6139672
[47,]  -0.2046700  -0.0783200  -1.0061500 -0.8874800   0.0717500  -0.7137600  -0.9890000  -0.3609800 -0.9907500  -1.2066800  -0.9963400
[48,]  -1.4576900  -0.1214200  -1.4622700 -1.4071500   0.0342300  -1.2171900  -1.3738600  -1.3683200 -1.4764900  -1.2158300  -1.2813700
[49,]   0.9566300   1.4909700   0.3876600  0.9906700   1.0424100   0.4957800   0.6935300   0.8964500  0.4581800   0.1681400   0.9393100
[50,] -11.1195560  -3.9267338  -7.4766097  1.2267498  -0.7629850  -4.9659693  -0.7440628  -2.2122772 -4.3768199  -5.8303785   2.5232953
```

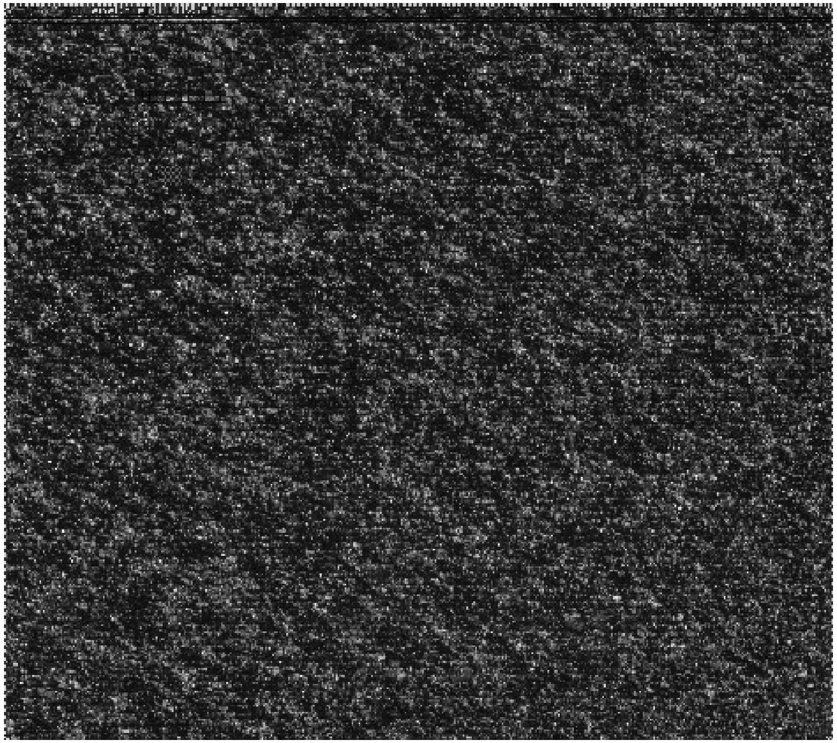
```
[10,] -0.21968299  -0.4279651   0.2644103  -5.7497198  -0.96908672  -1.7355809  -2.232554784  -1.1486723  -1.75152890  -3.65072248  -2.27827012
[11,] -0.08376162  -7.2164826  -3.8646045   0.7744070  -3.18557745  -0.8807509  -3.231808153  -1.0453582  -2.84251288  -3.43758030  -4.59190530
[12,] -1.08336656  -1.1468886  -1.2254498  -2.3613409   0.19293828  -3.3237007  -4.535195432  -4.6944589  -0.93266901  -0.55561048  -1.79152391
[13,]  0.04333000  -0.4637700   0.1275600  -0.3953500  -0.20215000  -0.0784500  -0.556270000  -0.0242100   0.16172000  -0.46966000  -0.57163000
[14,]  1.16542000   0.8624800   1.1604900   1.2394100   0.51927000   1.0825400   1.437250000   1.3304900   0.70357000   0.83907000   0.94630000
[15,] -0.29687000   0.2860200  -0.6962400  -1.1977900   0.19546000  -1.2171900   0.132900000   1.2014800  -0.02162000  -0.28167000  -0.98322000
[16,]  1.76249000   1.0755600   1.4620100   1.1607600   1.29921000   1.7220700   0.319450000   2.0086700   0.76309000   0.87675000   0.57842000
[17,]  0.46387000   0.2127100   0.4945500   0.5829000  -0.44349000   0.1769600   0.572750000   0.6968600   0.07154000   0.16814000   0.03714000
[18,] -17.27492074   3.9551562 -10.2644179  -2.9503713 -10.77896241 -12.0285402  -2.998205710  -1.0566138  -0.25583289 -10.35052138 -10.16008698
[19,]  0.25603000   1.3326000   0.9168600   0.2755000   1.44118000   0.6445200  -0.027010000  -0.3168000   0.89954000   1.27720000   1.58425000
[20,]  3.55160561 -13.6690025 -11.5670422  -5.4409270  -3.71877175  -1.8697248  -3.635858069   5.9623282 -12.31006506   0.89340328   0.44462681
[21,] -2.63154208  -4.4759312  -4.1105241  -0.4633036  -2.21023673  -6.8825294  -4.718002217  -2.6607977  -1.13554017  -0.23483606  -1.04912615
[22,]  0.04514000   0.8242800   0.2397200  -0.8767500   0.41242000  -0.5737500  -0.646060000  -0.0179700   0.23077000   0.73032000   0.16671000
[23,]  0.04152000  -0.1026900  -0.7454100  -0.2534600  -0.39054000  -0.0921600  -0.072240000   0.5560500  -0.20265000  -0.32607000   0.07030000
[24,]  1.31159000   0.8857100   1.3603100   1.0907600   0.55596000   1.3635100   1.811450000   1.6174700   1.60153000   0.88335000   0.99594000
[25,] -1.29186941   2.9366637  -2.0607812  -6.9160518   6.82540778  -6.8861316  -7.384511471  -2.5594320  -7.36637904   1.53987774  -6.55706133
[26,] -0.10796000  -1.3942000  -0.5476700   0.0781600   0.16770000  -0.3160300  -0.382770000  -1.1084000  -0.29503000  -1.04043000  -1.28137000
[27,]  0.98840000   0.3561800   0.6850200   1.3725500   1.52144000   0.8687400   1.826430000   1.2760600   0.44761000   0.36772000   0.54291000
[28,] -1.01269000  -0.9295900  -1.4622700  -0.7615200  -0.30942000  -1.0714500  -0.132050000   0.2195600  -1.47649000  -0.69566000  -0.98322000
[29,] -8.86074531  -5.5049797   2.2156872 -18.0968544  -3.51475893  -3.1800043  -4.113883182  -2.4237267 -12.09824840   0.76152772  -4.90184417
[30,] -1.16730000  -0.7804600  -1.4622700  -0.9373100  -1.03877000  -0.0077100   0.009790000   0.3747500  -1.02009000  -0.84890000  -0.64479000
[31,]  1.29401000   2.1756600  -0.2282000   0.5693100   0.80986000   0.8089600   0.814010000   1.0485800   0.69500000   0.96083000   0.37348000
[32,] -2.47282664  -4.1457904  -9.0093500  -5.2021172  -3.02712122  -2.2859246 -16.021147514  -2.5443746  -3.71706901   2.46944413   2.85273452
[33,] -0.10583000   0.7971200   0.5864200  -0.1159100   0.61404000   0.7417000  -0.326340000  -0.9572000   0.80544000   0.63325000   1.21005000
[34,] -15.17033400  -5.7929441  -5.2492469  -1.7173400  -8.08748532  -3.1660964 -13.906466777   2.9103923 -18.15936836 -19.38379581  12.71167007
[35,] -2.39545623   0.8817244  -3.9999577  -2.8785429   0.21497029 -12.4237349   3.382160545  -9.5392159   1.48635110  -5.31031316  -8.28417839
[36,] -9.79002539  -6.2150922   7.0833133 -23.1484993 -14.30620004 -12.9000765 [illegible]  -0.8595600
[37,] -0.55527000  -0.9295900  -0.5825200  -0.7336500  -0.38450000  -0.6767300  -0.556270000  -1.3683200  -1.08735000  -1.11981000  -0.60104000
[38,] -1.13866058  -0.9545288  -0.9879789  -0.7486456  -1.55059438  -1.1119910  -0.609846679  -0.5773113  -1.09804120  -1.07038147  -0.66548883
[39,] -4.31192770  -2.6357080  -5.2730756  -1.2860492  -3.50255805  -2.9157206  -3.089170656  -1.4861254  -3.18789106   3.04104876   1.37886654
[40,]  0.14612000   0.2806900   0.1115600   0.9573100   1.09138000  -0.2129600   0.368420000   0.0105800   0.53591000   0.79359000   0.50217000
[41,] -3.65834077  -4.0428742  -3.5751380  -2.9586557  -1.17516836   3.0683915  -0.531091769   1.5020583  -1.36423855   1.35179282  -2.34094331
[42,] -1.45769000  -0.6043900  -1.4622700  -0.2079400  -1.42668000  -0.5023100  -0.053890000  -1.3683200  -0.71418000  -1.21583000  -1.28137000
[43,] -1.45769000  -0.7500000  -1.4622700  -1.0020400  -0.57041000  -0.0977000  -1.373860000  -1.3683200   0.47207000  -0.41411000  -0.93255000
[44,]  0.41067000  -0.3158500   0.2818300  -0.2184600  -1.42668000   0.5944700  -0.000890000   0.1148000   0.62386000  -0.28836000   0.43537000
[45,] -0.92065000  -1.3942000  -0.1301900  -1.4071500   0.17595000  -0.5645100  -1.373860000  -1.3683200   0.20288000  -1.21583000   0.09439000
[46,] -0.72622000  -0.5900000  -1.2590800  -0.6802700  -0.46950000  -1.0714500  -0.583920000  -0.9867000  -0.66036000  -1.21583000  -0.17337000
[47,]  1.19892200  -0.0614334   2.2158800   0.4342396  -1.08822474   2.1175700   0.035228775  -0.8142278   2.21227000   0.85683943   1.03042010
[48,] -6.30569551  -3.0137140   1.6460885  -7.0379199 -15.13035310  -8.1522252   3.636102218  -5.0846871   2.47483148  -6.42108563  -7.71745619
[49,] -1.73054083  -2.1564257   0.3578752  -5.9754669  -0.06199337  -4.2852430  -0.013441783  -2.0239560   0.03406096   1.25826898  -0.47371781
[50,] -0.74307000  -0.9454500  -0.8099600  -0.8821000  -0.94055000  -0.4201100  -0.958570000  -0.5553400  -0.48433000  -1.21583000  -1.28137000
```

```
[10,]   1.3706866  -3.1815053   1.0945986 -2.2518298  -0.3100496  -0.6686597  -0.6540535  -1.9287335 -3.2379697  -1.2578805  -3.7475174
[11,]   1.0482100   1.3511300   1.2707900  0.9512500   1.3523800   1.3059000   1.1234800   0.8849800  1.2081000   1.5833800   1.1002200
[12,]  -0.2838600  -0.2799200  -0.0740500 -0.0607000  -0.1874500  -0.0489800  -0.9407800   0.2707600  0.0189300  -0.1034100   0.5968100
[13,]  -0.0826900  -0.4920800   0.1461000  1.0789000  -1.4266800   0.6920100  -0.3478800   1.0291000 -0.7605600  -1.2158300   0.0005400
[14,]   1.2655600   0.6419300   1.4654800  1.8207700   2.0073600   1.4918500   1.4782200  -0.2940800  1.9542600   1.2704300   1.2622500
[15,]   1.0793800   0.8870500   1.2652700  1.3917500   1.5701100   1.3536700   1.3709700   1.0549900  1.1460900   1.3019800   1.0526800
[16,]  -0.5766680  -1.7953233   0.4123687 -1.3000260  -0.3914956  -0.7176529  -1.9670290  -1.2006734 -0.5781646  -0.1262439  -1.9169129
[17,]   0.7748900   0.4785400   0.6538900  0.5344200   0.6830700   0.6679600   0.6442800   0.7885300  1.0363900   0.7424500   0.9729600
[18,]   0.2745500  -0.0585100   0.3185400  0.2388900  -0.3607500  -0.0649300   0.1958400   0.4509100 -0.1332400   0.2465000   0.0005400
[19,]  -1.7879147  -1.0191236  -1.5552526 -3.5924412  -1.7454766  -0.7730166  -0.3718211  -0.5105483 -1.2629439  -2.9315562  -0.6414949
[20,]   0.5260500   0.3957500   0.6546800  0.5320500   0.2394300   0.6618500   0.5165000   0.6978100  0.2828800   0.3952900   0.3588100
[21,]  -1.4576900  -0.0863700  -1.4622700 -1.4071500  -1.4266800  -1.2171900  -1.3738600  -1.3683200 -1.1429700   0.6344900  -0.8615000
[22,]  -9.6719893   2.2601265   7.7628864 -3.4866526   4.3207356  -7.2708633 -13.5159651   2.1101495  1.2602673  -6.0691419 -12.5633527
[23,]  -1.6251345  -1.7400089  -0.8286638 -1.8216632  -1.4516043  -1.0365167  -0.4819336  -0.8287533 -1.4291445  -2.1557932  -0.3753139
[24,]   0.2773600  -0.8494900  -0.1880900  0.7294600   0.8723400   0.2823900   0.5251000  -0.1515900  0.4077000  -1.2158300  -0.2254300
[25,]   0.9831800   1.3916500   1.4639100  0.8695400   0.9272300   1.4933500   1.7161000   1.3445300  1.3880900   1.4714300   1.1528100
[26,]  -0.4879006   2.3522517   3.9431369 -0.7394428  -4.5868684  -4.9465152  -4.8069718  -3.3983413 -4.1992253   0.4114557  -2.9567431
[27,] -12.3719528   1.2538366  -6.3306388 -8.3371014   0.8558988   0.7665416  -2.1466177  -7.3020149 -1.8094012   2.9542900   0.9333962
[28,]  -1.0775500  -1.3942000  -0.8933400 -0.9148300  -1.4266800  -0.5972500  -0.7218800   0.1395300 -0.8057400  -1.2158300  -1.2813700
[29,]  -5.4636983  -8.7262323 -11.6165240 -9.5747269  -2.9136981  -5.7593651  -5.6418661  -5.9271453 -7.5683810  -7.1324658  -7.1897210
[30,]   0.8047400   1.1665100   0.9457400  0.7891100   0.2988400   1.2077600   0.9021200   0.9256000  0.7943100   1.1191800   1.1347800
[31,]   1.6280900   1.3590700   1.7875200  1.6325900   1.5195200   1.8001000   1.7636900   1.9085500  1.5812000   1.4344900   1.6826600
[32,] -15.9177707  -4.4406363   4.7738056 -7.3396969  -3.9618396 -10.7848498 -10.1883121 -10.6706668  3.8301984 -15.6259941 -12.0914823
[33,]   0.6881200   1.2516600   0.7484200  0.9140500   1.4083300   0.8687400   0.6728900   0.4837400  1.0766700   1.1625200   1.1341500
[34,]   0.3385600   0.5253400  -0.0596600 -0.8249000   0.8082000   0.9212000  -0.7312200  -0.7411500  0.8827000   0.8231600   1.1157400
[35,]  -0.7821400   0.0807700  -0.3072600 -0.7021500   0.0662200  -0.4280100  -0.2200700   0.0004800  0.3340100  -0.7439600   0.4983800
  -0.8595600  -0.0839000  -0.9927200 -0.5058000   0.0757300  -0.4963000
[37,]   0.2327100   1.0574200   0.7632300  0.8612200   1.4853800   0.7473100   0.3884400   0.1653000  1.1433600   1.0985900   1.1872300
[38,]  -0.1015800  -0.4637700   0.1461000 -0.0856000  -0.7132200  -0.2647300   0.2773900  -0.2664600 -0.0805300  -0.2425100  -0.2511800
[39,]  -0.0480300   0.4828000  -0.3608500 -0.2984100   0.6687000  -0.2256400  -0.0538900  -0.5666200  0.4772200   0.1803300   0.0307900
[40,]   0.6457300  -0.3533100   0.4244100 -0.1821700   0.3117000   0.1769600   0.3567800  -0.1983400  0.0130500  -1.2158300   0.4621500
[41,] -13.2327629  -7.4784300  -2.3736338  3.6926116  -0.1132352  -9.2827358 -10.1963165 -10.2116792 -4.4631231 -18.6168573  -6.8622131
[42,]  -1.1448447   1.6734572   0.2015480 -0.2543426  -3.2946844  -2.1751983  -0.1855370  -0.6094323 -1.8925225  -0.2569482  -0.6987012
[43,]  -1.0331500   7.8505726  -3.4200628 -0.5226154 -11.0591912 -13.9737873  -7.0726104 -18.2617373  5.1637078   7.8912059   1.2117123
[44,]  -5.2197951 -11.8997017  -8.4754373 -6.5459542 -11.5095411 -11.7889469  -5.3662040  -4.9429571 -6.7202350   1.1376922  -8.4510359
[45,]  -0.9708282  -6.1642629   0.8067003 -1.4437932  -1.6959041  -0.1198638  -1.4731559  -1.3784706 -4.1129785  -0.1576396  -2.8251738
[46,]  -2.1051862   0.6287777  -3.4271863  0.7024729  -1.8750100  -0.6556041   8.5340851  -2.6672457 -1.5931683  -2.3232690  -2.6139672
[47,]  -0.2046700  -0.0783200  -1.0061500 -0.8874800   0.0717500  -0.7137600  -0.9890000  -0.3609800 -0.9907500  -1.2066800  -0.9963400
[48,]  -1.4576900  -0.1214200  -1.4622700 -1.4071500   0.0342300  -1.2171900  -1.3738600  -1.3683200 -1.4764900  -1.2158300  -1.2813700
[49,]   0.9566300   1.4909700   0.3876600  0.9906700   1.0424100   0.4957800   0.6935300   0.8964500  0.4581800   0.1681400   0.9393100
[50,] -11.1195560  -3.9267338  -7.4766097  1.2267498  -0.7629850  -4.9659693  -0.7440628  -2.2122772 -4.3768199  -5.8303785   2.5232953
```

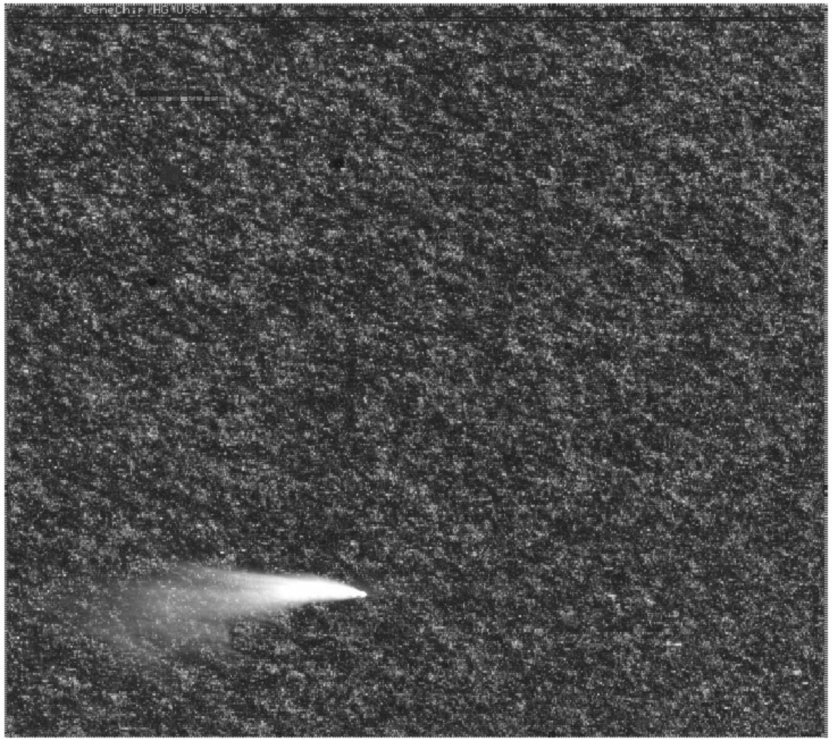
Quality control

Plot the raw image of the array

```
> image(…)
```



Nothing special

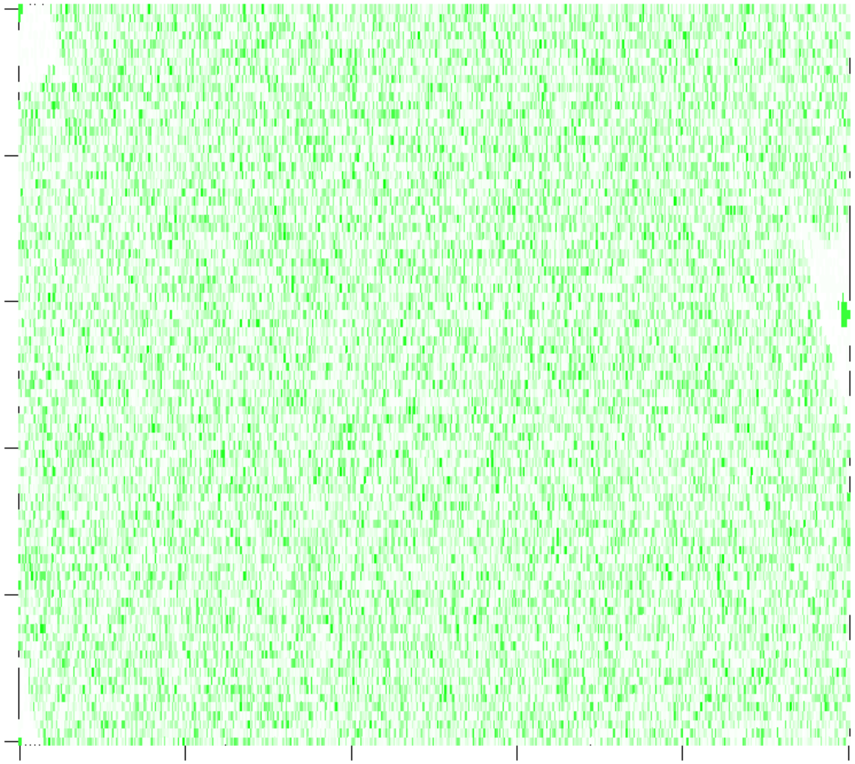


…

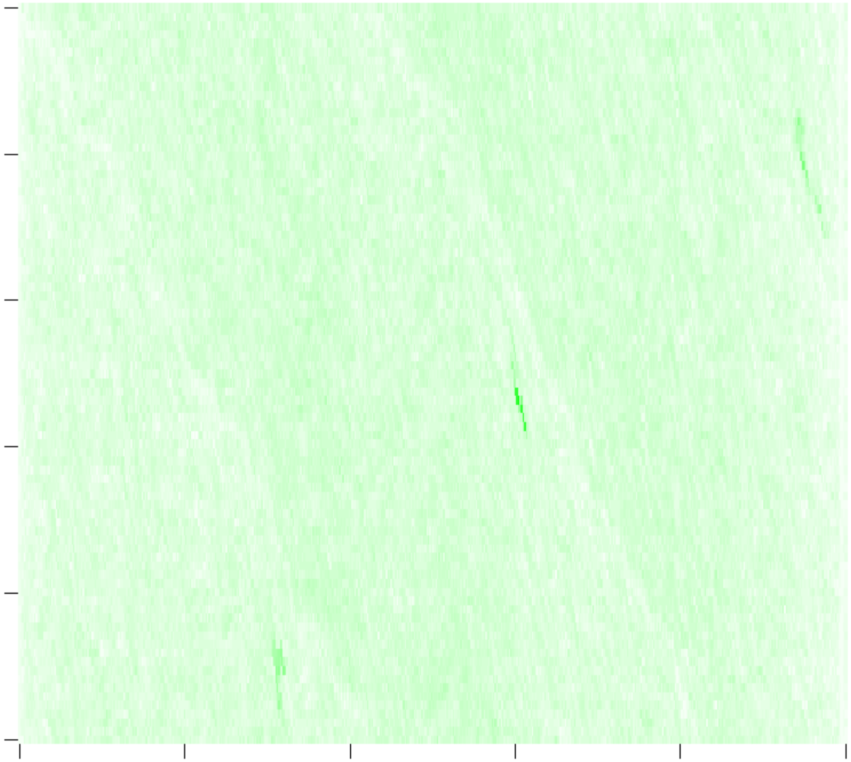
Quality control

Plot image of fore- and background signal

```
> image(…)
```



Foreground signal

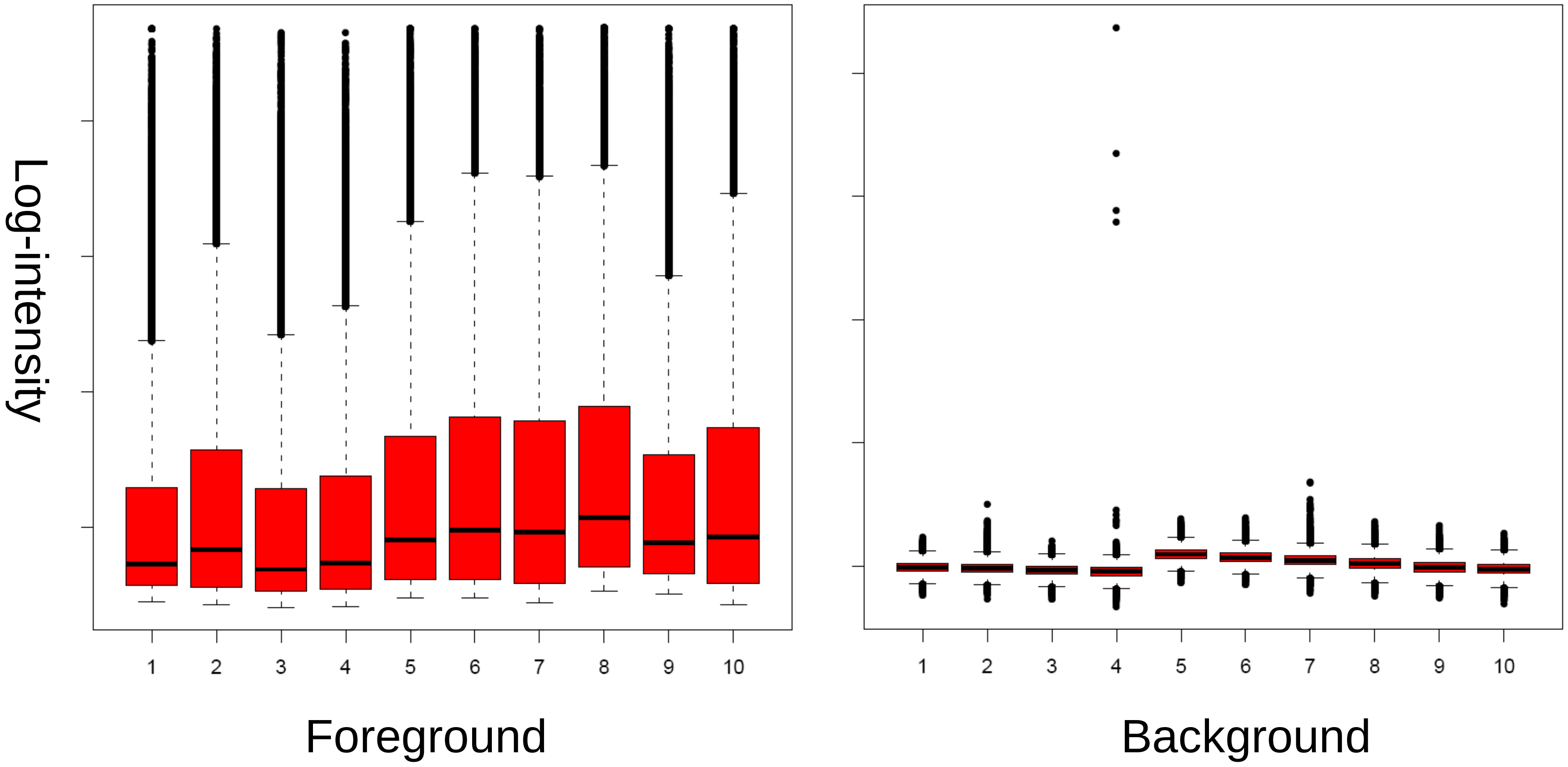


Background signal

Quality control

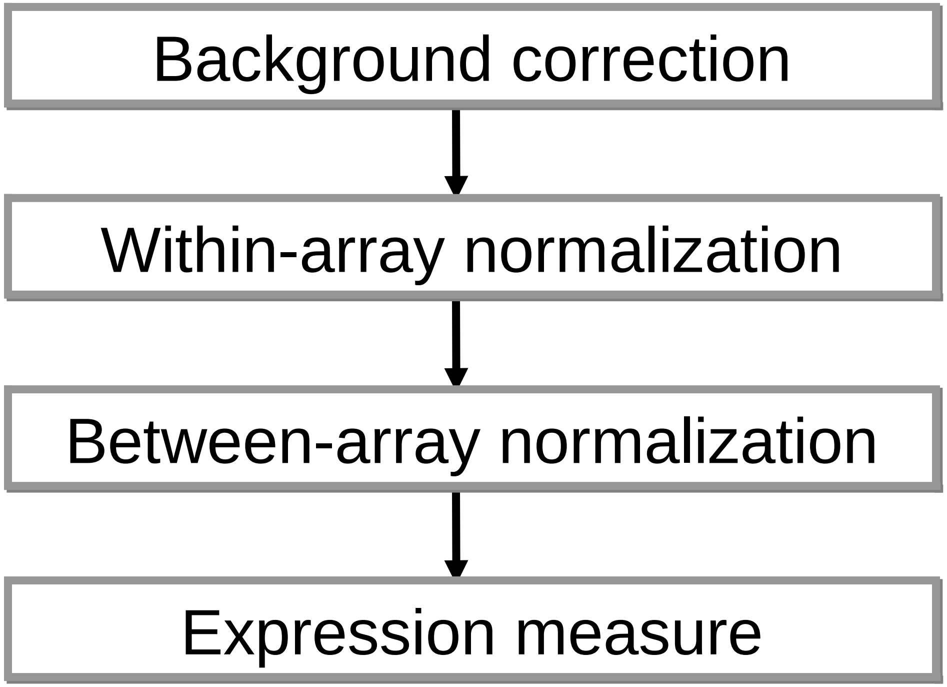
Generate boxplots of fore- and background signal

```
> boxplot(…)
```



Preprocessing

Before the statistical analysis of interest, the gene expression measurements (intensities) undergo several preprocessing steps.



Background correction

Within-array normalization

Between-array normalization

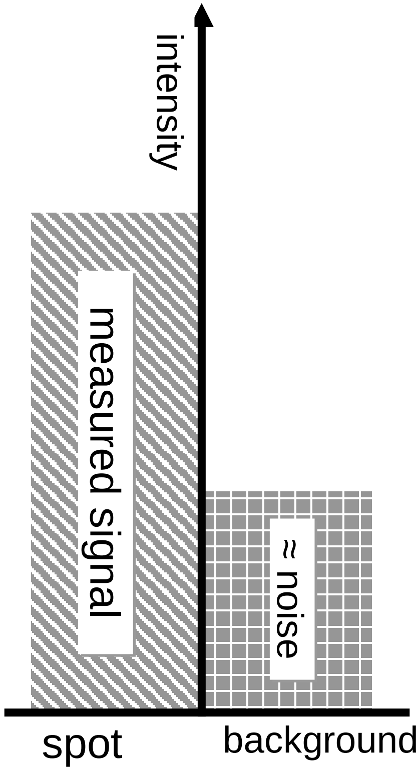
Expression measure

Background correction

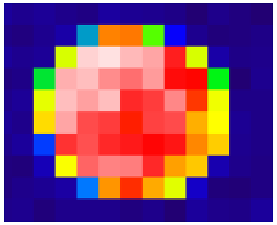
Background intensity

The background of the microarray may have a non-zero intensity.

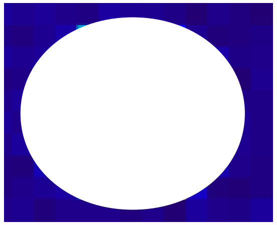
Hence, a feature's intensity may include a contribution not specifically due to the hybrization of the target to the probe.



spot

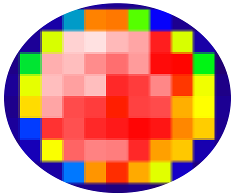


spot surrounding



nonzero intensity: not necessary representative of background noise inside feature area.

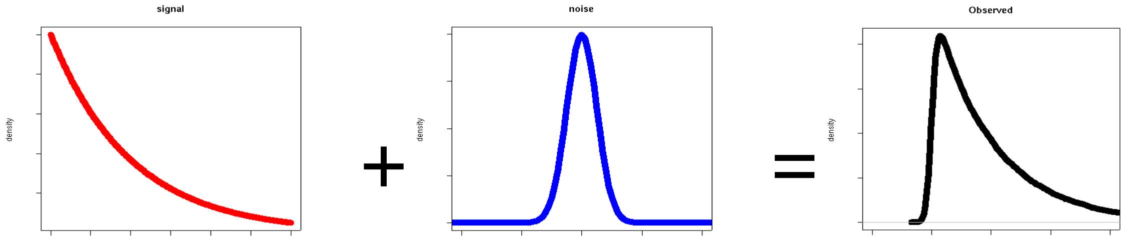
signal



contains non-zero intensity background noise

Background correction

Many signal-noise models view the observed log-intensities as a combination of true signal and background noise.



Signal + Noise = Observed

The ‘signal + background’ model for the intensities:

$$Y_{ij} = S_{ij} + BG_{ij}$$

$\Rightarrow$ Y_{ij} is the intensity of sample *i* and feature *j*.

$\Rightarrow$ S_{ij} and BG_{ij} are independent random variables.

$\Rightarrow$ $BG_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$

$\Rightarrow$ $S_{ij} \sim \exp(\alpha_i)$

Background correction

Estimation of μ_i :

- Fit a density to the Y_{ij}, using a kernel density estimator.
- Estimate μ_i by the mode of the density.

Estimation of α_i and σ_i :

$$\hat{\alpha}_i = \sum_{j=1}^{p} (Y_{ij} - \hat{\mu}_i) I_{\{Y_{ij} \geq \hat{\mu}_i\}} / \sum_{j=1}^{p} I_{\{Y_{ij} \geq \hat{\mu}_i\}}$$

$$\hat{\sigma}_i^2 = 2 \sum_{j=1}^{p} (Y_{ij} - \hat{\mu}_i)^2 I_{\{Y_{ij} \geq \hat{\mu}_i\}} / \left(\sum_{j=1}^{p} I_{\{Y_{ij} \geq \hat{\mu}_i\}} - 1 \right)$$

The background corrected intensity is $B(Y_{ij}) = \mathbb{E}(S_{ij} \mid Y_{ij})$ with

$$\mathbb{E}(S_{ij} \mid Y_{ij}) = \int_0^\infty sP(Y_{ij} \mid S = s) P(S = s) / P(Y_{ij}) ds$$

and

$$f_Y(Y_{ij}) = \int_{-\infty}^{\infty} f(Y_{ij} - z) f_{BG}(z) dz$$

Background correction

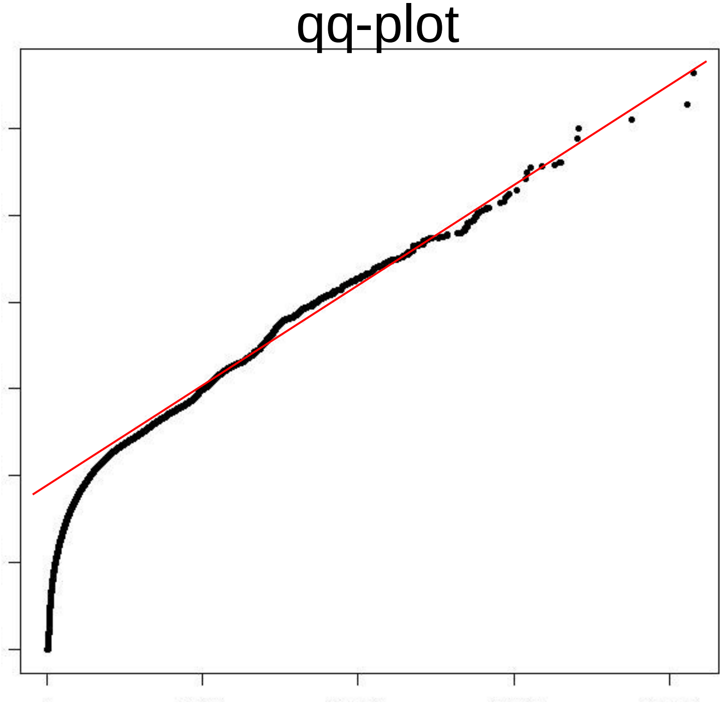
Important

There is no a priori justification for the presented ‘signal + background’ model (or any of its competitors):

its usefulness must prove itself in application!

For instance,
check distributional
assumptions.

qq-plot



Normalization

Motivation for normalization

Normalization is required to correct for experimental artifacts while preserving the true biological signal.

Normalization balances intensities

→ between dyes, and

→ between hybridizations

in order to allow comparison of gene expression across hybridizations.

- Dyes: within-array normalization.
- Hybridizations: between-array normalization.

Normalization

Conceptually, normalization adjusts intensities relative to intensities of reference genes whose levels are assumed to be constants between samples.

A set of genes that are to function as reference genes in the normalization must be chosen.

Genes for normalization

1. All genes on the array
2. Constantly expressed genes
3. Controls
4. Rank invariant genes

Normalization

1. All genes on the array

All genes on the array are used in normalization.

This is sensible when:

a) only a relatively small proportion of the genes will vary significantly in expression between mRNA-samples, or

b) there is symmetry in the expression levels of the up/down-regulated genes.

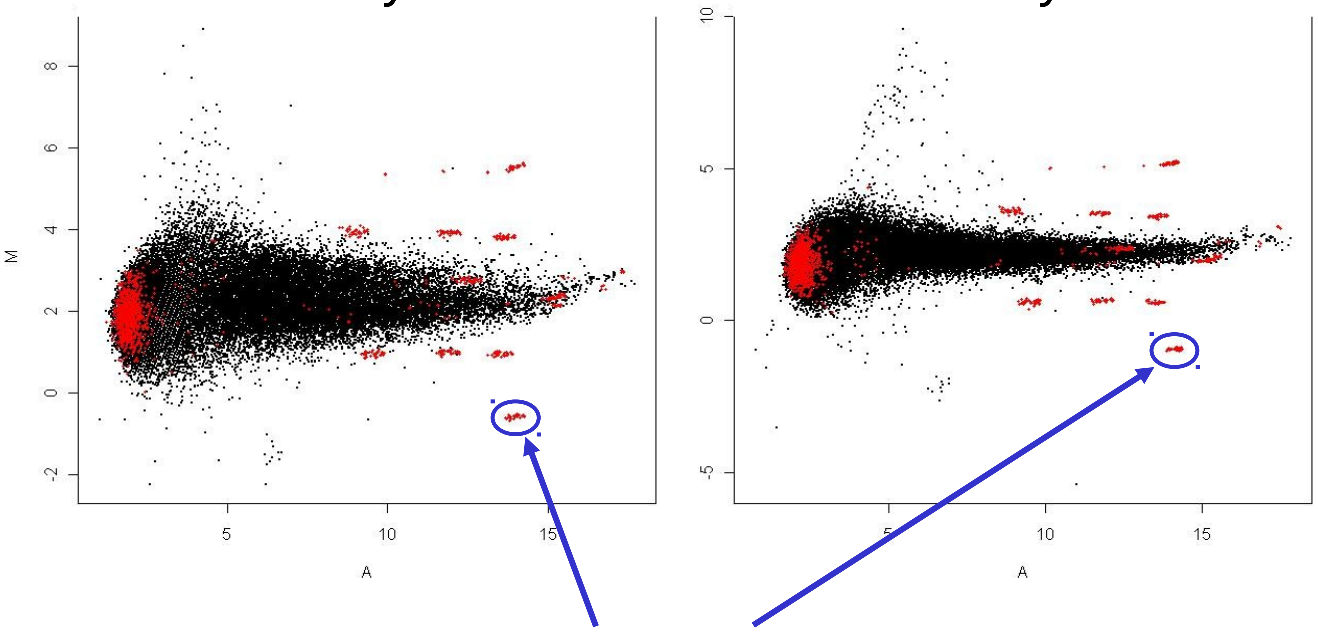
Normalization

3. Controls

Spiked-in controls are synthetic DNA sequences (complementary to oligonucleotides on the array) and included in the mRNA samples at equal amount and should have equal intensities across hybridization.

array 1

array 2



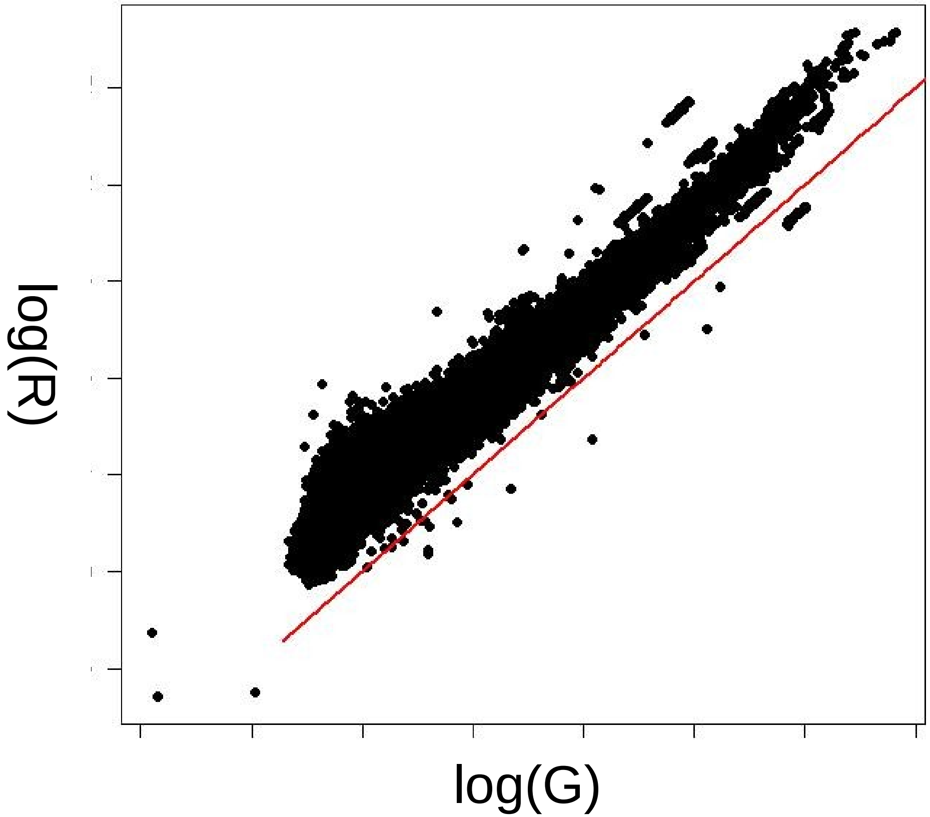
spiked-in controls

Normalization

Within-array normalization aims

→ to balance intensities of the two dyes, as well as

→ to elimate other systematic differences due to unequal experimental conditions.



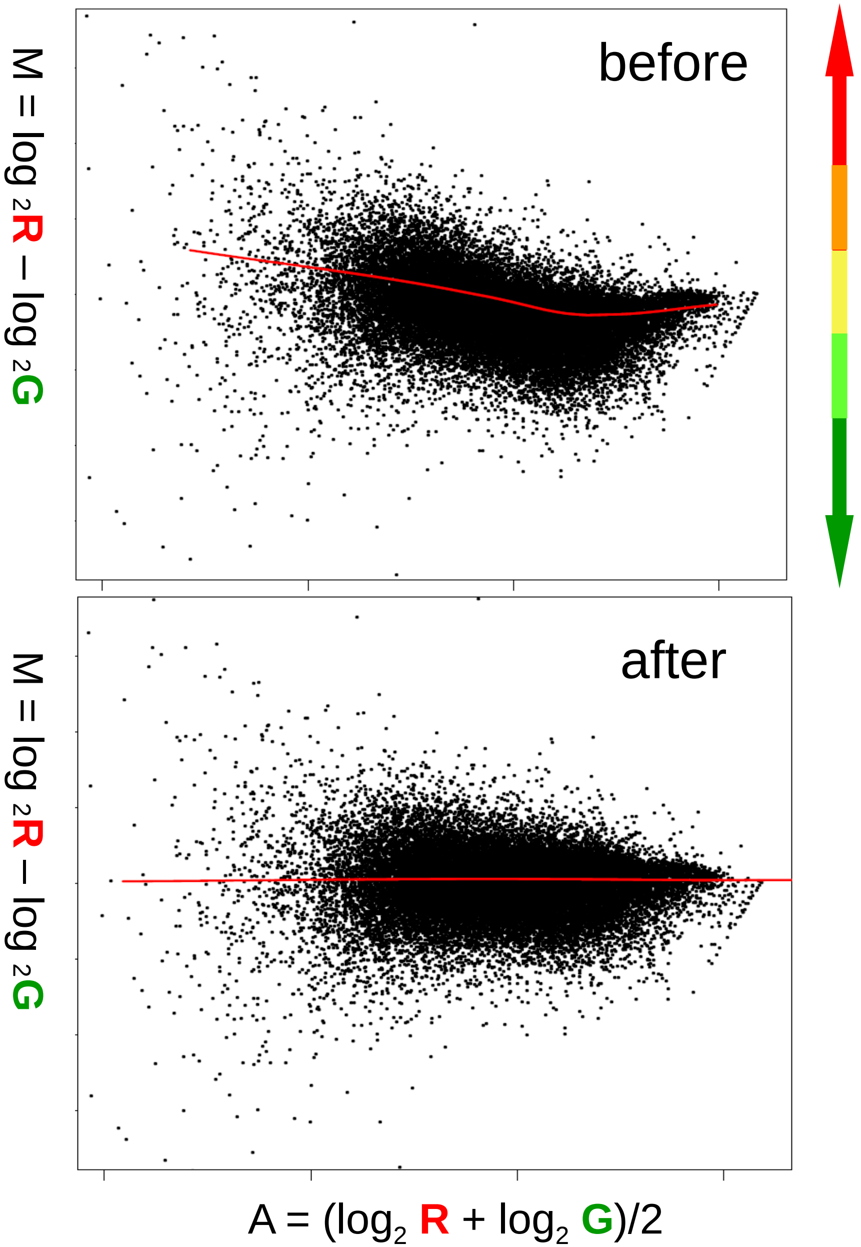
Systematic deviations from the line log(G)=log(R) indicate a dye-effect that is to be eliminated by normalization.

Normalization

MA plot

Within-array normalization uses the MA-plot to identify artifacts and detect intensity-dependent patterns in log-ratio's M_j.

Statistically, *within-array normalization* subtracts a function $g(\bullet)$ from the individual intensity log-ratio's M_j. The function $g(\bullet)$ is computed per array.



Normalization

Operationalization of between-array normalization

A transformation of the individual intensities values (or log-ratio's) such that the intensities are comparable across arrays.

→ A transformation is constructed for each hybridization.

→ The functional form of the normalizing transformation is determined by the type of normalization.

Discuss: scale and quantile normalization

Normalization

Scale normalization

Assume log-ratios from array *i* follow $\mathcal{N}(0, a_i^2 \sigma^2)$.

The scale factors a_i are robustly estimated by:

$$\hat{a}_i = \frac{\text{mad}_i}{\sqrt[n]{\prod_{i=1}^{n} \text{mad}_i}}$$

where mad_i is the median absolute deviation of array *i*:

$$\text{mad} = \text{median}_i\{|X_i - \text{median}_j(X_j)|\}$$

Scale normalization is then achieved by dividing the log-ratios by the estimate of a_i.

Normalization

Quantile normalization aims to make the distribution of probe intensities the same across arrays. This operationalization is motivated by the assumption that the amount of mRNA in each sample is roughly the same.

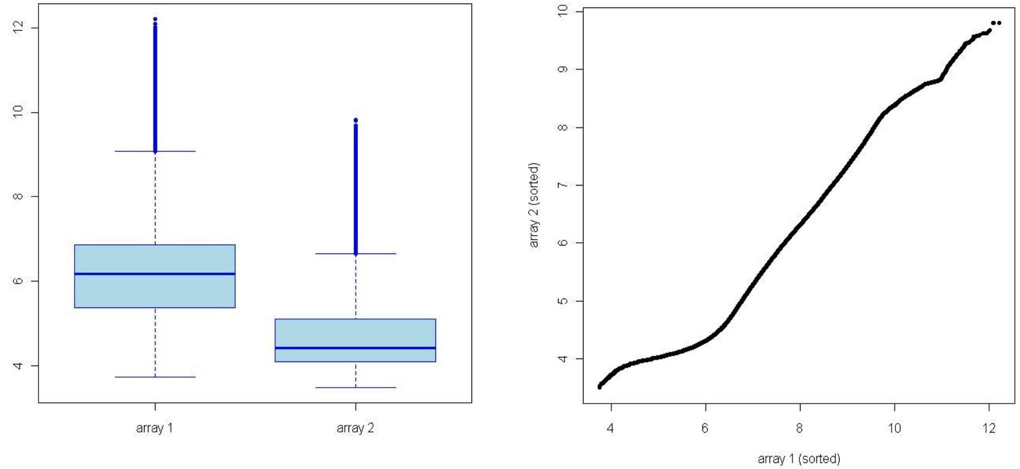
It transforms the data from all arrays such that the transformed data follow the *n*-dimensional identity line in the *n*-dimensional qq-plot.

Rationale

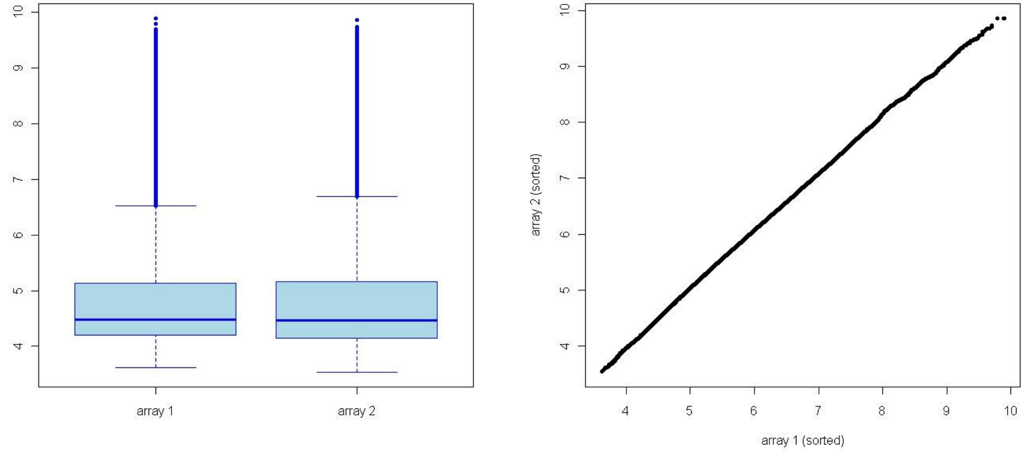
The quantiles of two identical distribution line up on the diagonal of a qq-plot. This suggests that two datasets could be given the same distribution by equalling their quantiles.

Normalization

Different distributions ...



… same distribution



Expression measure

Operationalization of expression

An expression measure is a number reflecting the amount of RNA in the sample.

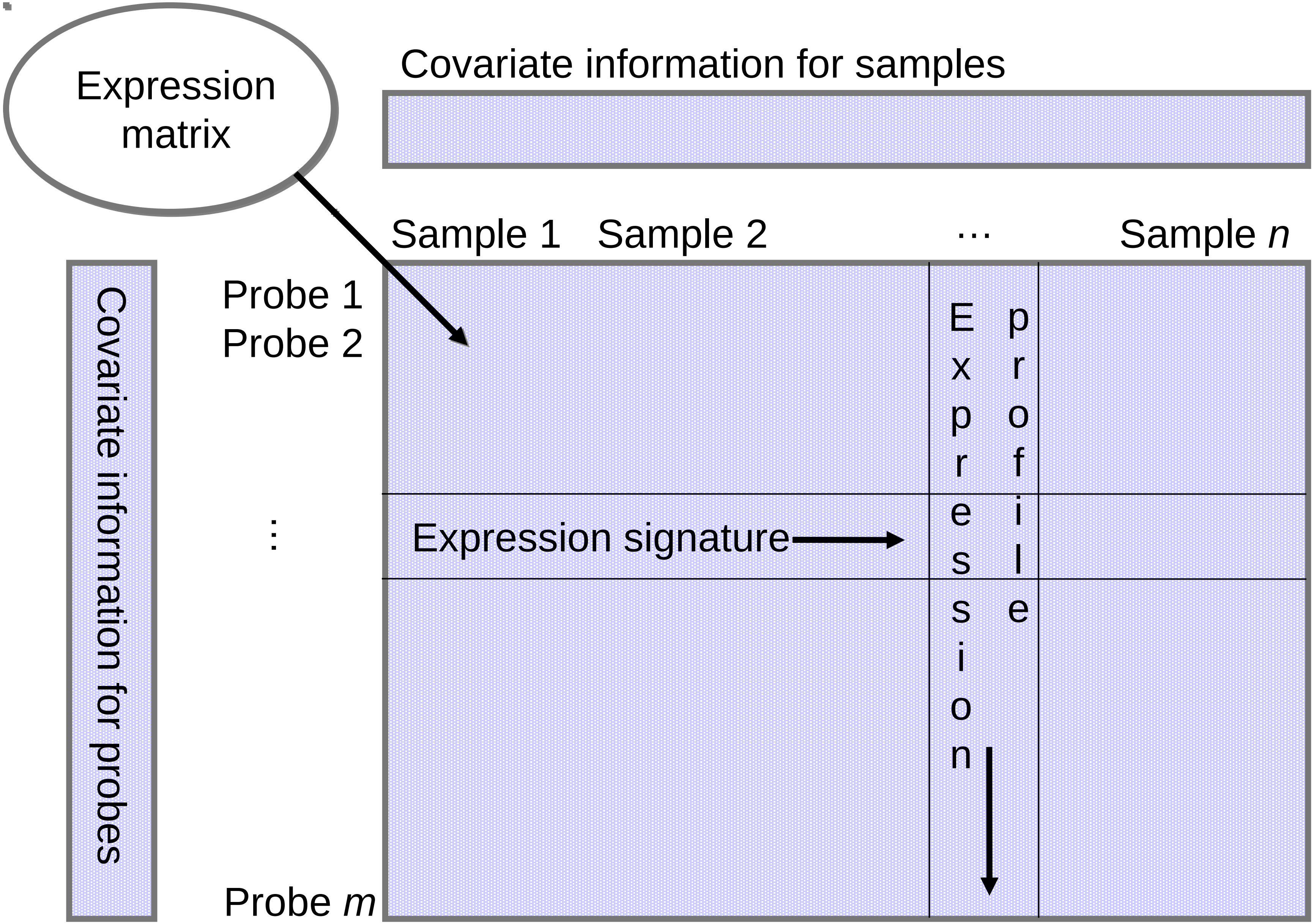
For dual channel arrays:

- the expression measure is simply the log-ratio's M_j.

For the Affymetrix single channel array:

- an expression measure is determined by summarizing the probe level data of a probe set (set of features interrogating the same gene) into one number.
  (not discussed)

Expression measure



Expression matrix

Covariate information for samples

Sample 1 Sample 2 … Sample *n*

Covariate information for probes

Probe 1

Probe 2

⋮

Probe *m*

Expression signature

Expression profile

Rubbish?

“Microarrays are the closest thing to fraud we accept in science.”

-- *????, ????*

- Inherently noisy.
- Many sources of variation.
- Many preprocessing steps, with lots of arbitrary choices.

Rubbish?

Affymetrix spike-in experiment

→ 14 gene groups are spiked-in at varies concentrations in accordance with a latin square design.

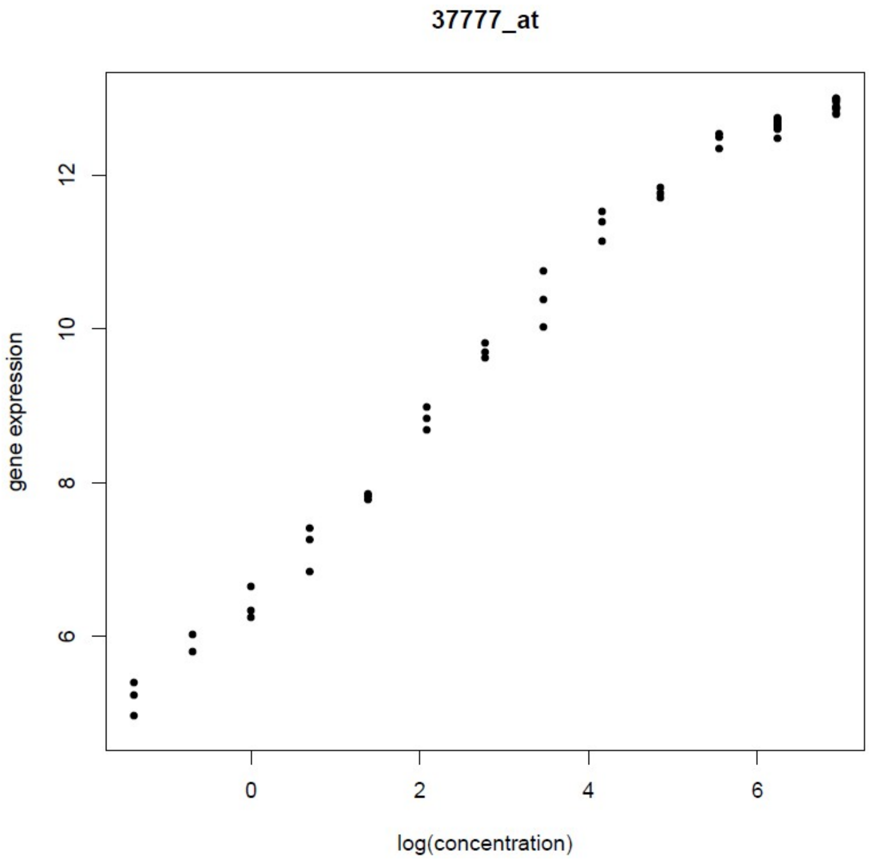
→ Each hybridization has been replicated at least three times.

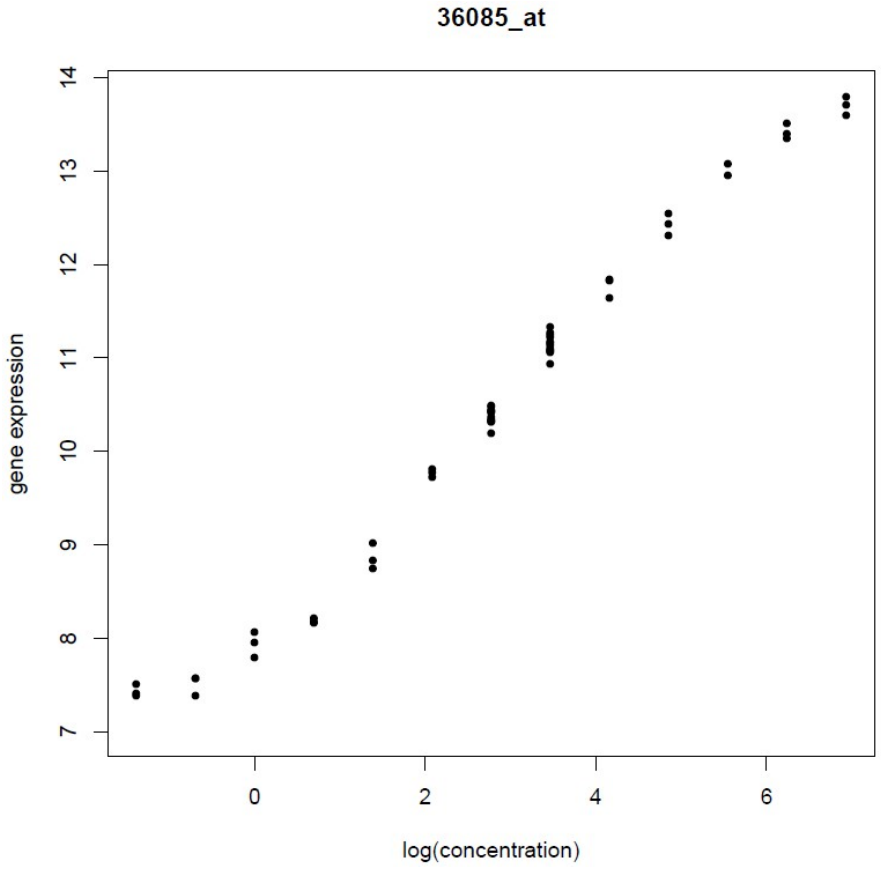
→ In total 59 hybridization.

→ Array type: HG-U95.

A proof of principle?

Rubbish?


37777_at
gene expression
6
8
10
12
0
2
4
6
log(concentration)

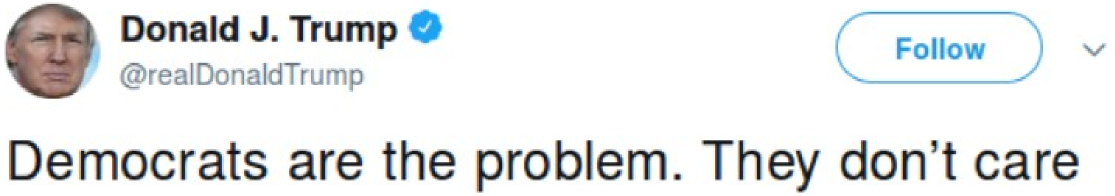

36085_at
gene expression
7
8
9
10
11
12
13
14
0
2
4
6
log(concentration)

Other data have similar issues

Twitter data

E.g. can one identify one's political preference on the basis of his/her tweets?

Sometimes easy:

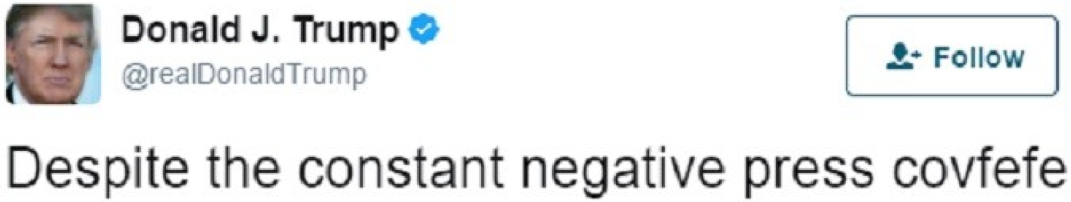


Donald J. Trump

@realDonaldTrump

Democrats are the problem. They don’t care

but meaning not always obvious:



Donald J. Trump

@realDonaldTrump

Despite the constant negative press covfefe

Other data have similar issues

Twitter data

Harvesting: which tweets to select?

Decide upon:

→ Time period of tweets.

→ Original tweet only?

→ Include retweets?

→ Include replies?

→ Which language?

→ Users' geographical location.

→ Include meta-data like user profile?

Other data have similar issues

Twitter data

Preprocessing issues:

→ URLs, @, #, emoticons, and other symbols,

→ Spelling:

  → 1000, 1,000, 1000.00, 1,000.00, or thousand, or

  → colour or color,

→ Synonyms:

  → loud or noisy (in e.g. a Tripadvisor review),

→ Acronyms:

  → POTUS, LOL, BFF.

→ Remove low frequency words?

→ Remove stop words like “and”?

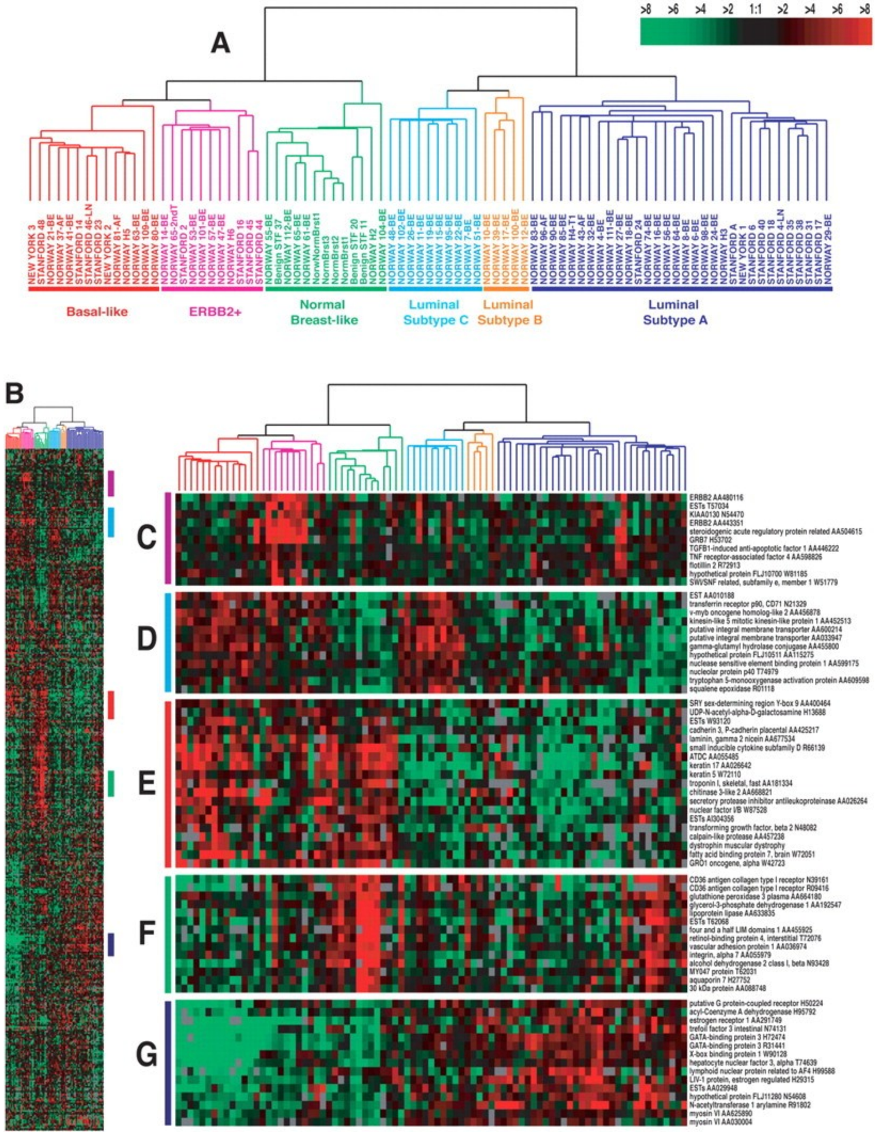
→ Combine tweets?

An example:
the big promise

An example

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

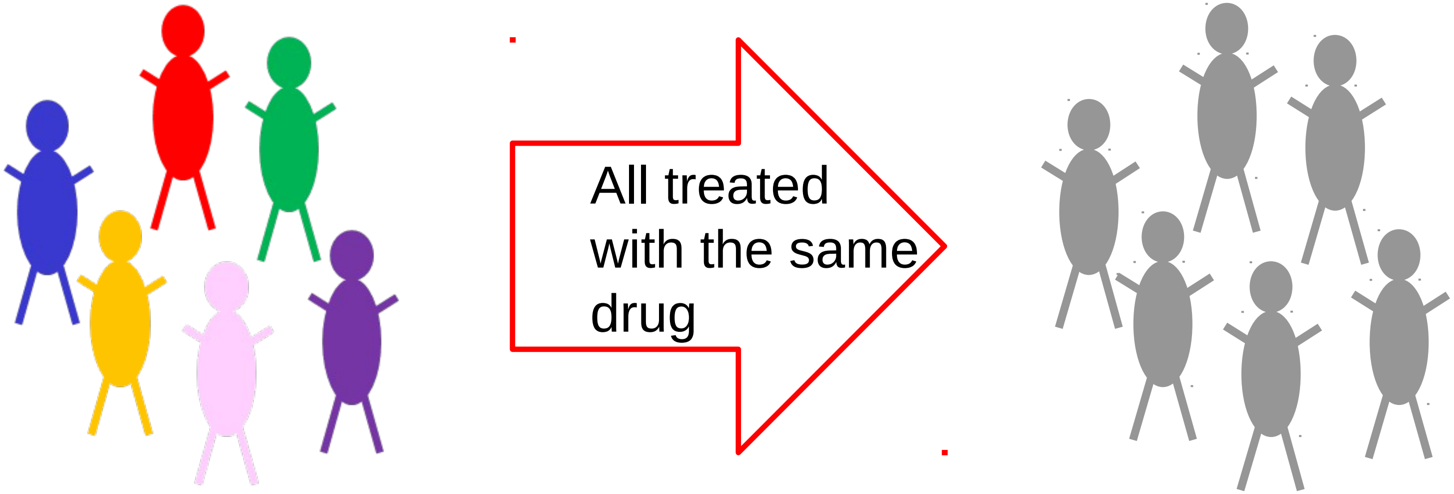


Using 78 breast cancer profiles, five subtypes are distinguished:

- Basal
- ERBB2
- Luminal A
- Luminal B
- Normal

An example

Traditional medicine



All treated with the same drug

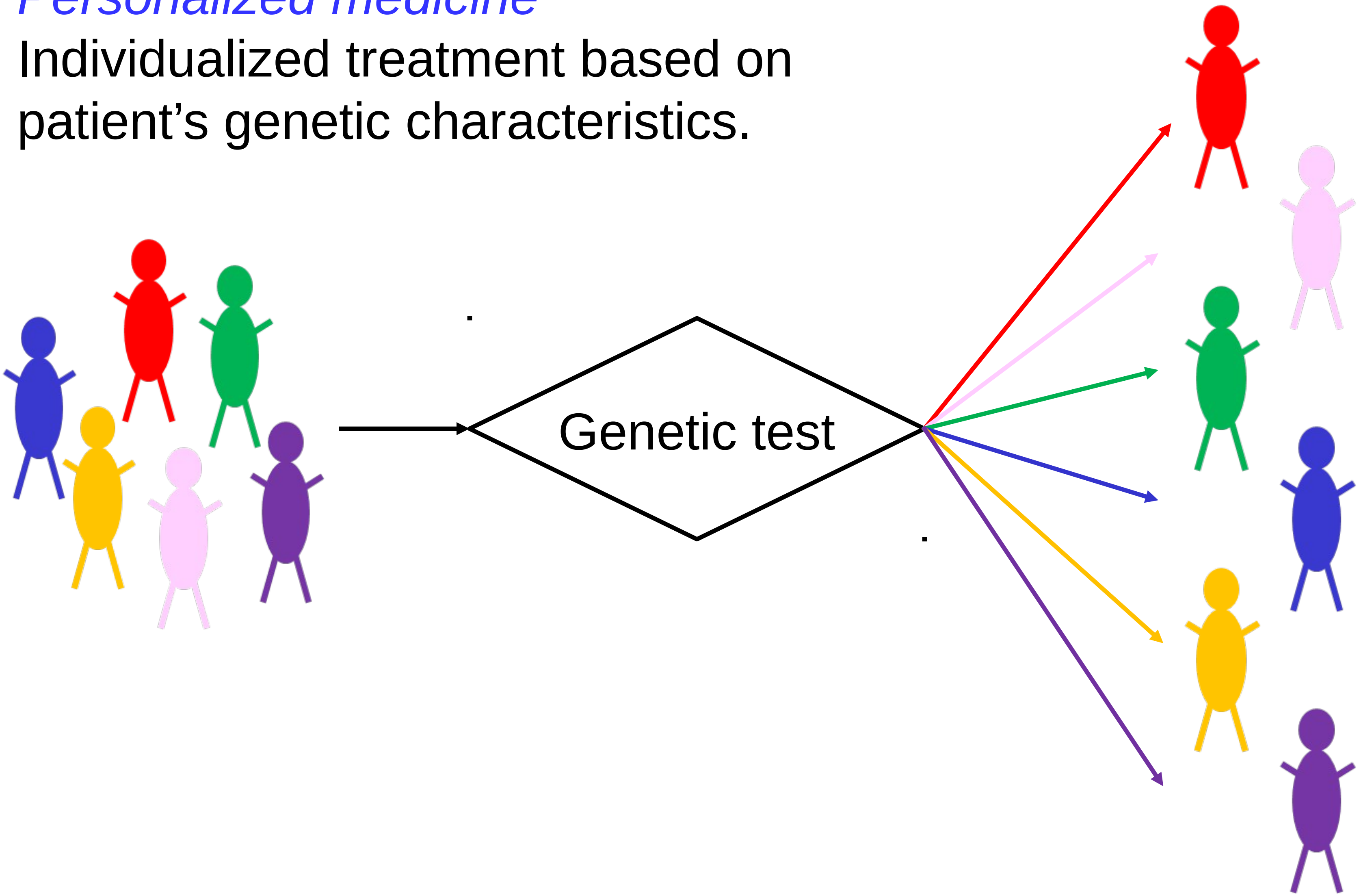
Standard treatment may not be beneficial to everyone.

Subgrouping of breast cancers suggest patients from different subgroups may need different therapy.

An example

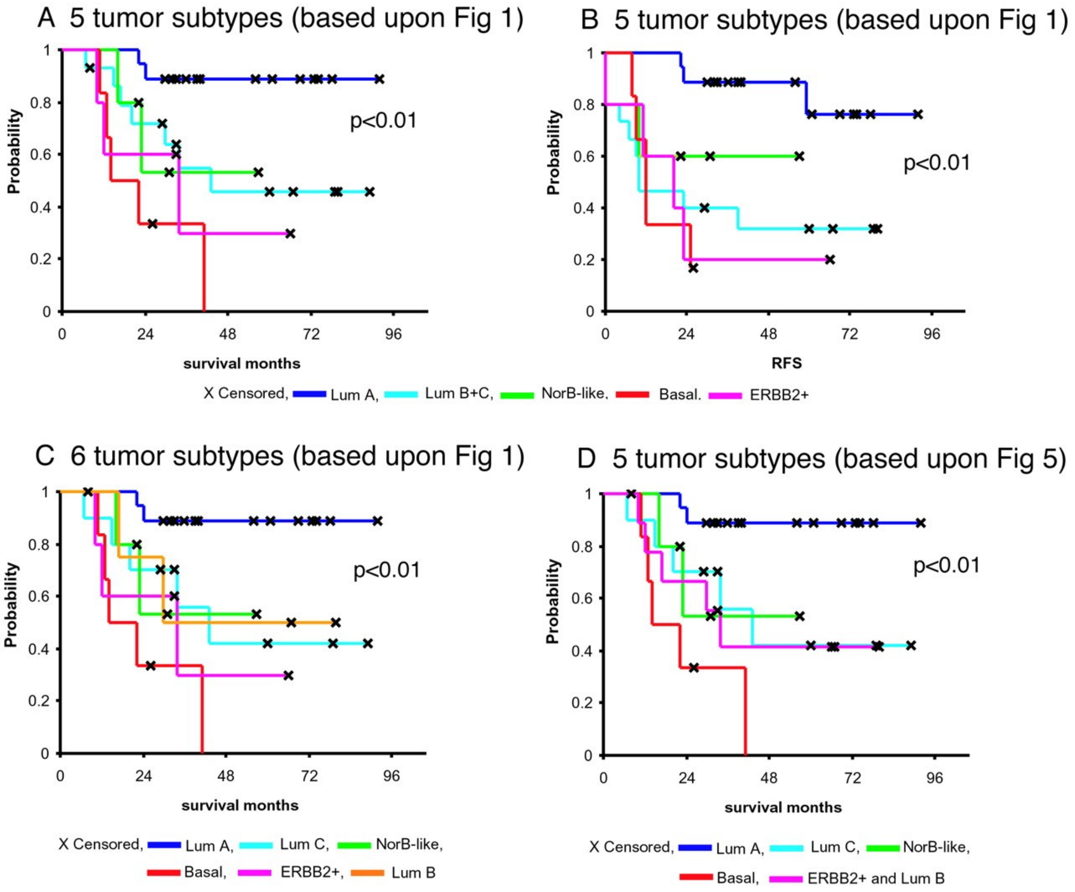
Personalized medicine

Individualized treatment based on patient’s genetic characteristics.



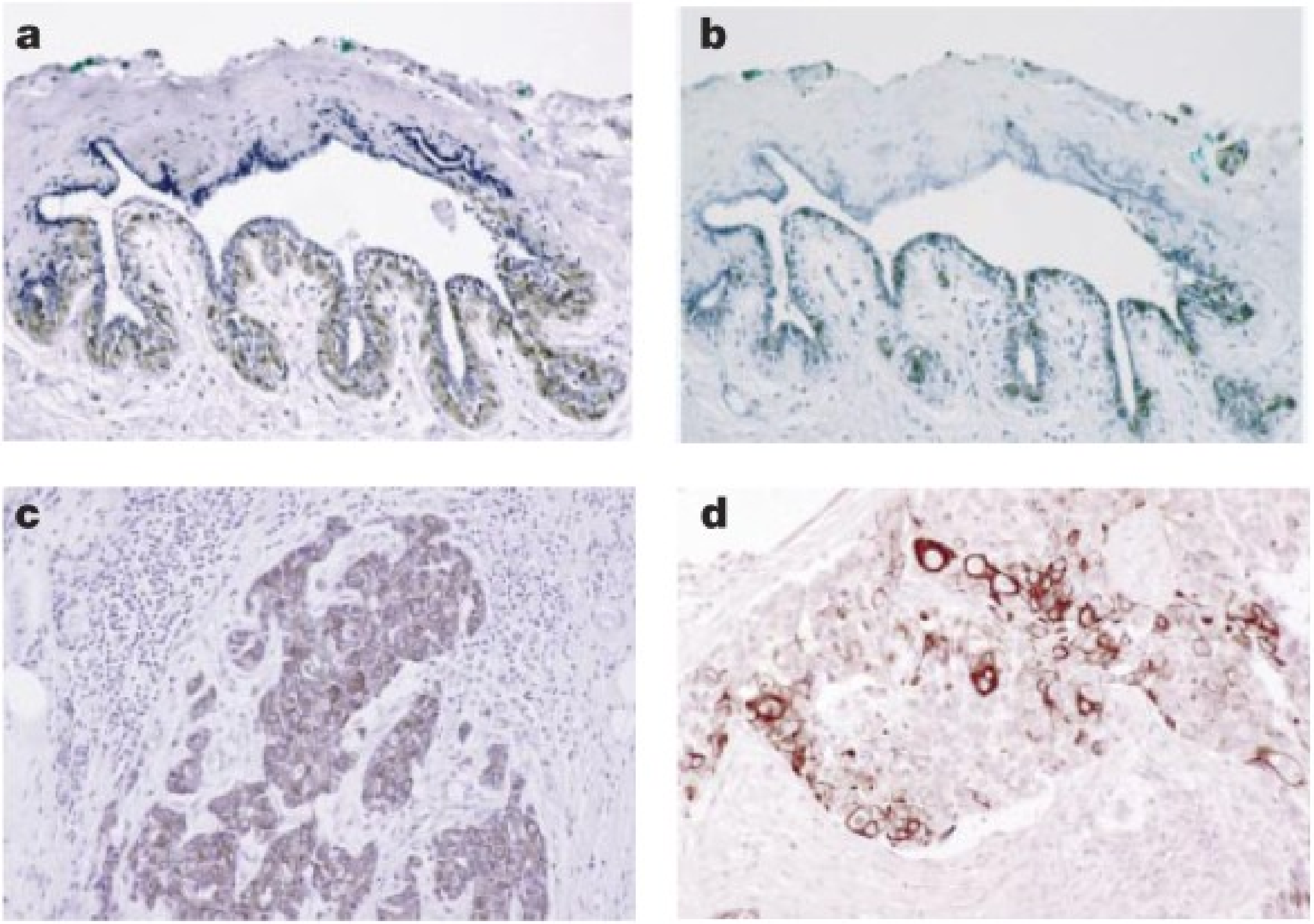
An example

Why do people believe these breast cancer subtypes?
1) Subtypes exhibit different clinical outcome.



An example

Why do people believe these breast cancer subtypes?
2) Exhibit different morphology.

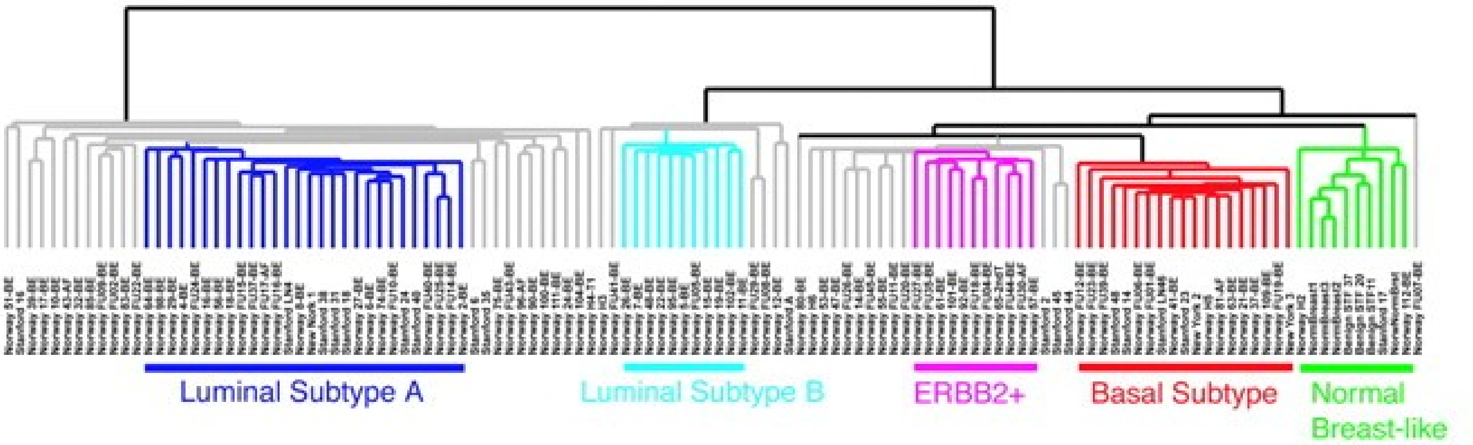


An example

Why do people believe these breast cancer subtypes?
3) Subtypes have been confirmed.

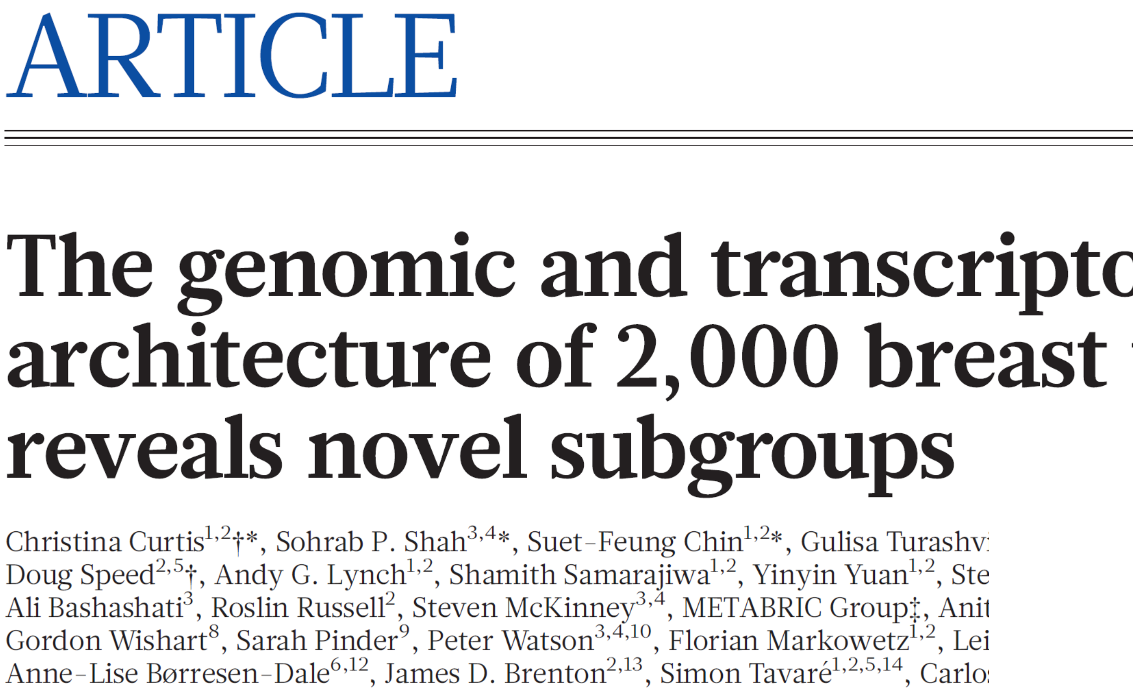
Repeated observation of breast tumor subtypes in independent gene expression data sets

Therese Sørlie[*], Robert Tibshirani[†], Joel Parker[‡], Trevor Hastie[§], J. S. Marron[¶], Andrew Nobel[¶], Shibing Deng[||], Hilde Johnsen[**], Robert Pesich[*], Stephanie Geisler[††], Janos Demeter[*], Charles M. Perou[‡,‡‡], Per E. Lønning[††], Patrick O. Brown[§§], Anne-Lise Børresen-Dale[**], and David Botstein[*¶¶]



An example

Medio 2012, the story continues …



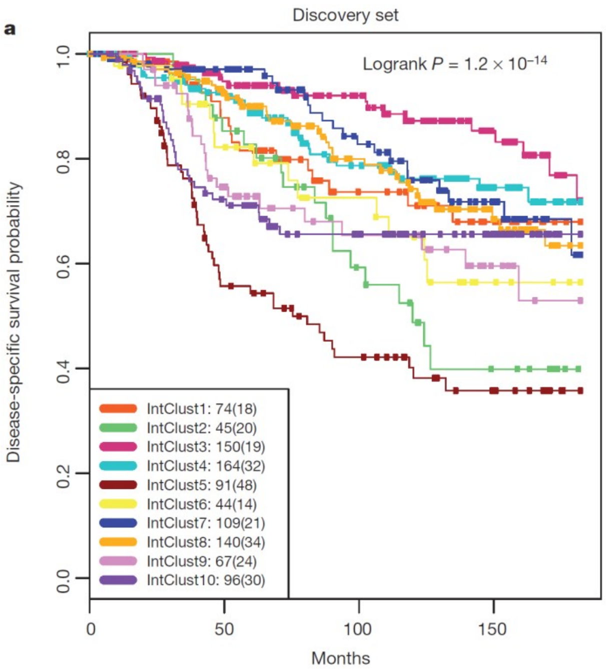
ARTICLE

doi:10.1038/nature10983

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis[1,2]†*, Sohrab P. Shah[3,4]*, Suet-Feung Chin[1,2]*, Gulisa Turashv
Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Ste
Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anit
Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Lei
Anne-Lise Børresen-Dale[6,12], James D. Brenton[2,13], Simon Tavaré[1,2,5,14], Carlo

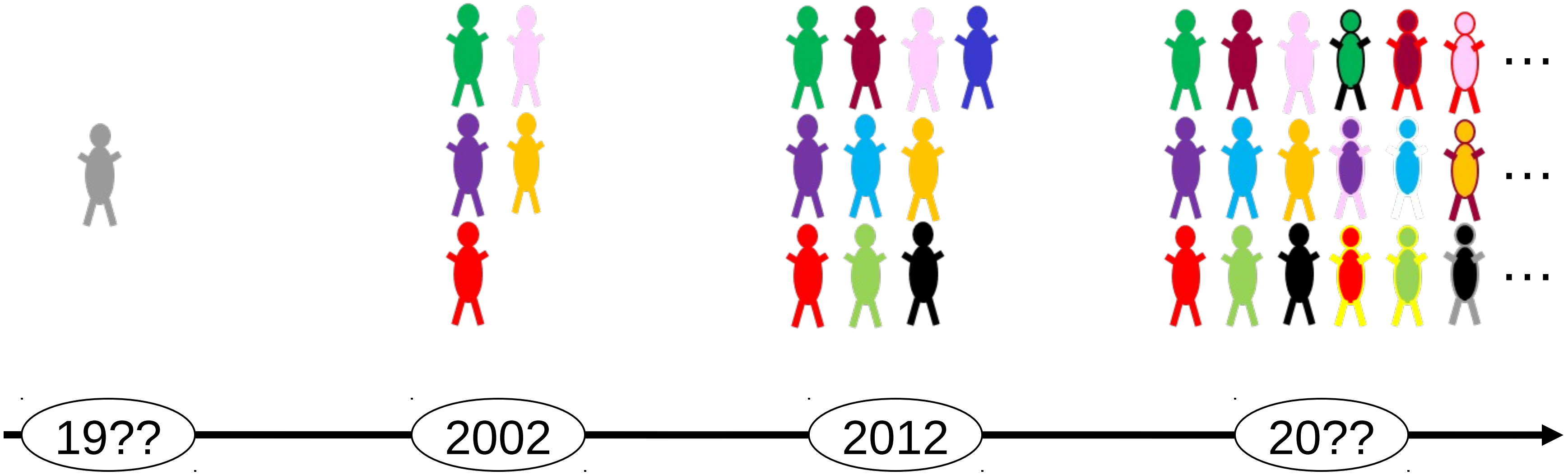
Inclusion of more molecular information suggests the existence of 10 subgroups.



An example

How many subgroups really exist?

Genetically, everybody is unique. Thus …



8 billion?

…

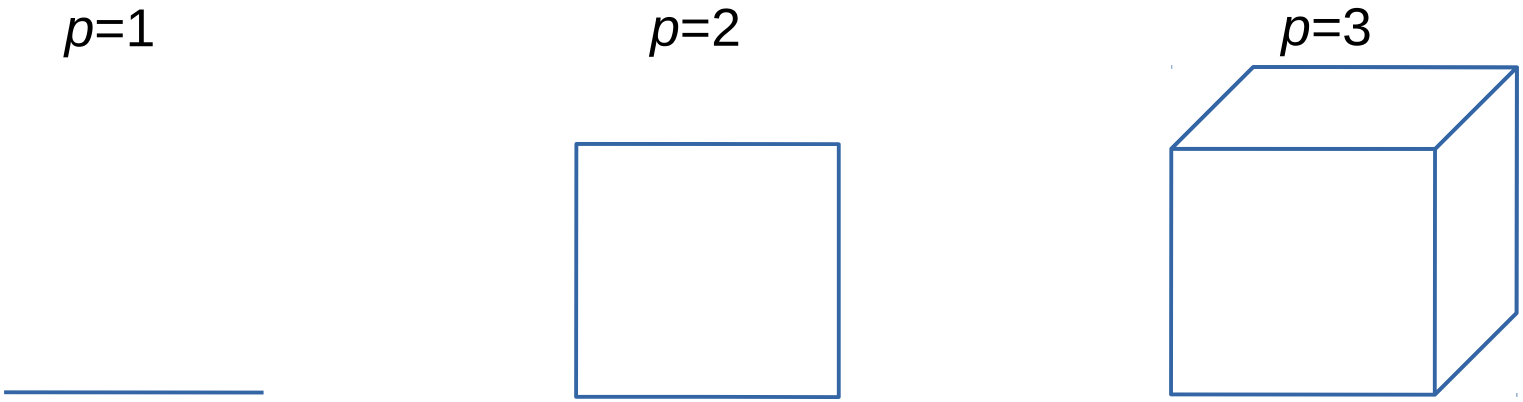
personalized medicine
to the max?

The curse of dimensionality

Curse of dimensionality

i) The high-dimensional space is enormous and data points are isolated.

Unit cubes

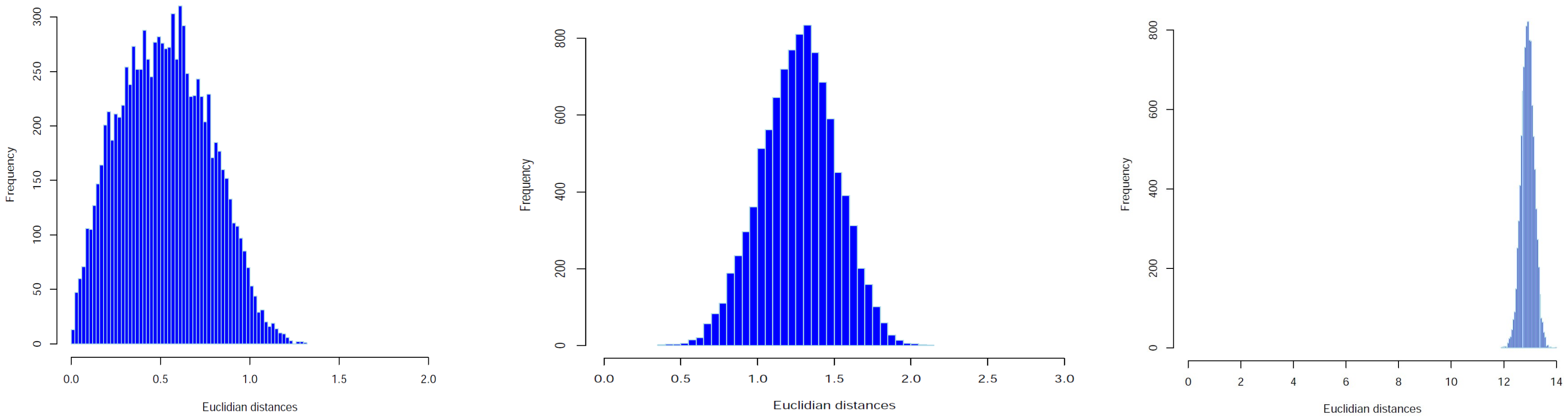


Maximum distances

$$\sqrt{1^2} = 1 \qquad \sqrt{1^2 + 1^2} = \sqrt{2} \qquad \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$$

Curse of dimensionality

Histograms of Euclidean distances between random vectors from the *p*-dimensional unit cube.



p=2 *p*=10 *p*=1000

Distances grow with *p* and (relatively) more homogeneous.

Question

What does this do to the “nearest neighbor” concept?

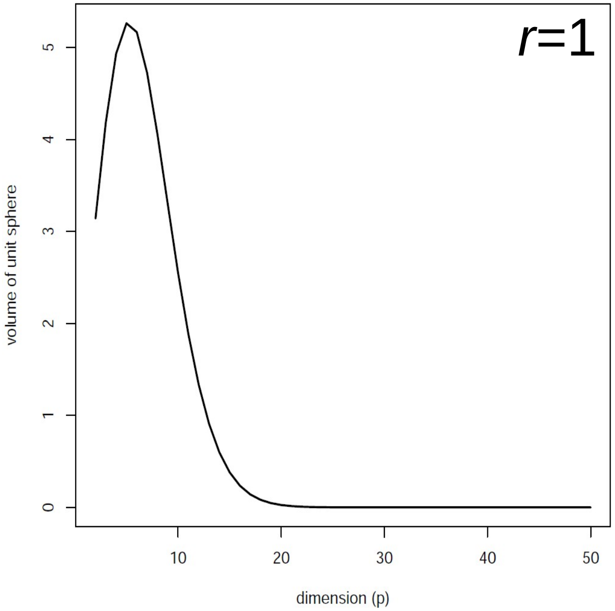
Curse of dimensionality

ii) Volume of unit balls vanishes as *p* increases.

The volume of a sphere with radius *r* in a *p*-dimensions:

$$V_p(r) = [\Gamma(p/2+1)]^{-1}\pi^{p/2}r^p$$

For $p = 20$, $V_p(1) = 1.73 \times 10^{-13}$.



Question

What are the consequences for the sample size?

Curse of dimensionality

A well-designed experiment provides good coverage of the design space.

→ Design space: unit cube $[0,1]^p$.
→ Distribute *n* points s.t. union of unit balls around the points encompass the unit cube.
→ Calculate required *n* for varying dimensions *p.*

p	*n*
20	39
30	43630
50	5.7×10^{12}
100	4.2×10^{39}
150	1.28×10^{72}

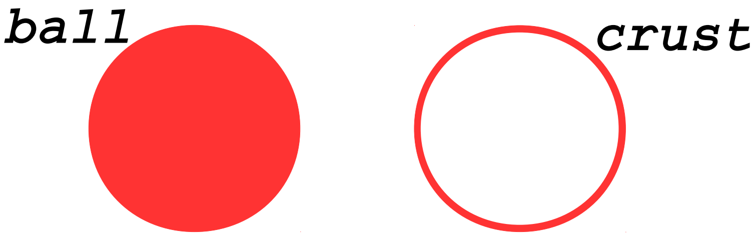
Curse of dimensionality

iii) Volume of unit balls concentrates around the crust.

Denote *p*-dimensional ball with radius *r* by: $B_p(0, r)$.

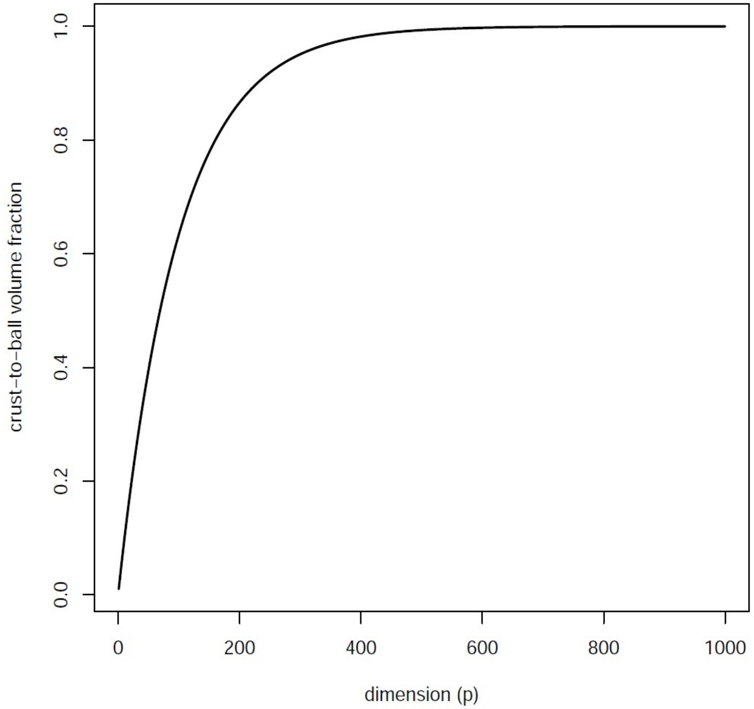
Define the “crust” by:

$$C_p(r) = B_p(0, r) \setminus B_p(0, 0.99\, r).$$



Then, the crust-to-ball volume is:

$$\text{Vol}[C_p(r)] \,/\, \text{Vol}[B_p(0, r)]$$

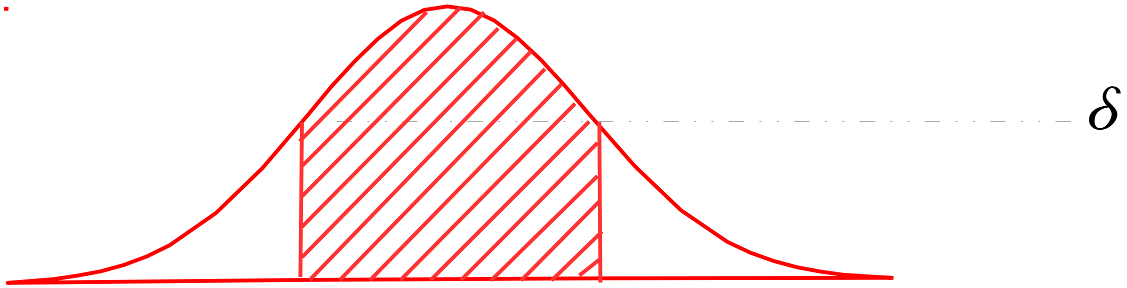


Curse of dimensionality

The univariate normal distribution concentrates most mass around its mean and has thin tails.

This is reversed for large *p*:

$$P[\exp(-\tfrac{1}{2}\|\mathbf{X}\|_2^2) \geq \delta] \leq (\delta 2^{p/2})^{-1}$$



Consequence
Rare events may not be so rare.

This lecture series

Issues: multiplicity

A comparison of the expression levels of gene *A* between two groups:

```
        Welch Two Sample t-test

data:  group 1 and group 2
t = -8.6449, df = 17.284, p-value = 1.099e-07
alternative hypothesis: true difference in means is not equal to 0
```

- This is a rather small *p*-value.
- Getting such a small *p*-value is unlikely.
- Is it still unlikely if we acknowledge that the gene is one of 40000 on the microarray?

Slide courtesy of Jelle Goeman

Issues: multiplicity

The multiplicity problem

Each individual test has a specified type I error probability. This probability of committing a type I error increases with the number of tests.

The probability of at least one false positive finding in m tests is given by:

$$1 - (1 - P(\sim H_0 \mid H_0))^m$$

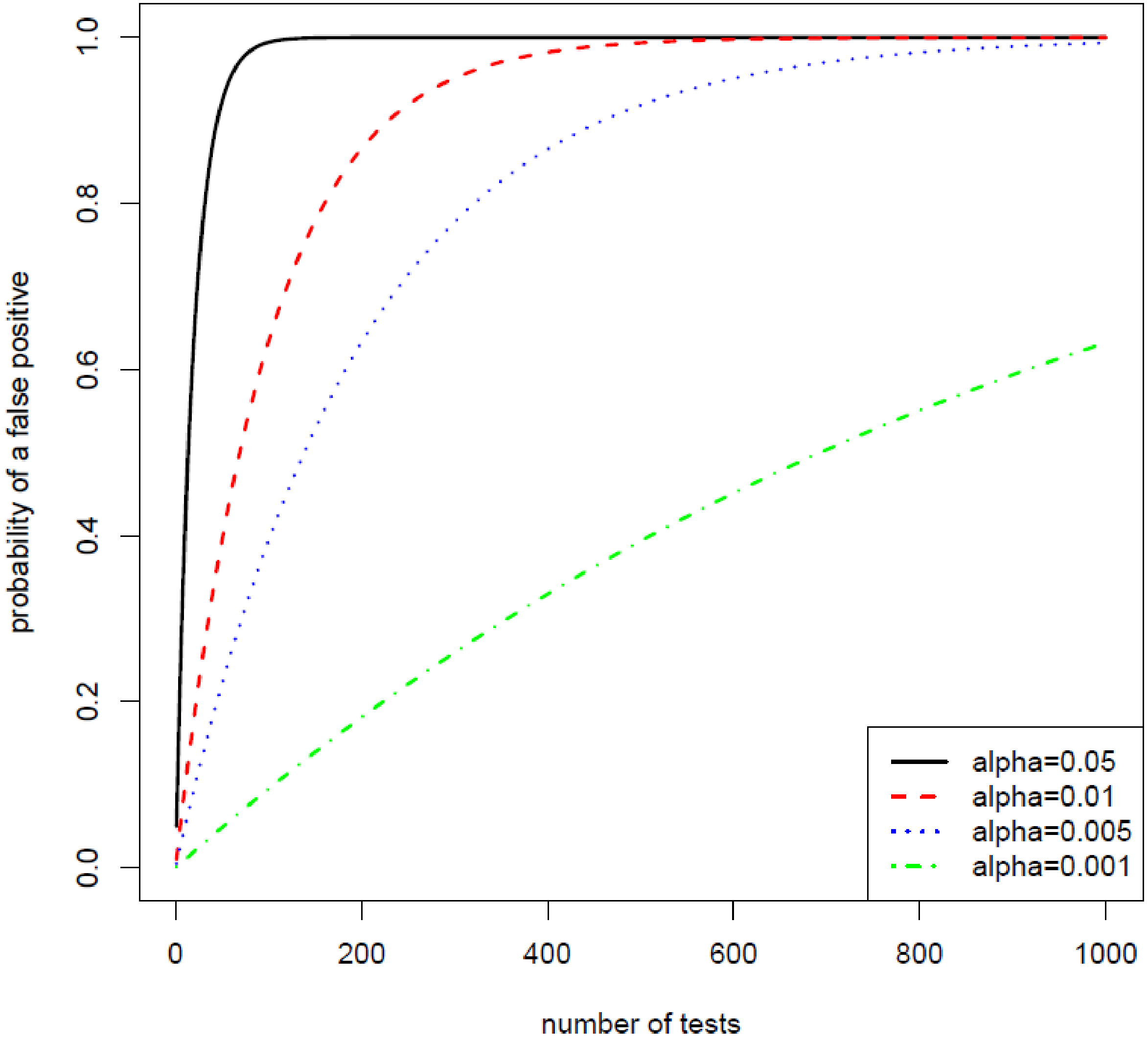
$$=$$

$$1 - (1 - \alpha)^m$$

m	$1 - (1 - P(\sim H_0 \mid H_0))^m$
1	0.0500
2	0.0975
5	0.2262
10	0.4013
100	0.9941

Issues: multiplicity

Decreasing the rejection level reduces the probability of a false positive.



Issues: multiplicity

Problem

→ many traits, many tests,

→ large number of false positives.

Multiple testing

→ generalization of type I error,

→ control of this generalized type I error,

→ control of number of false positives.

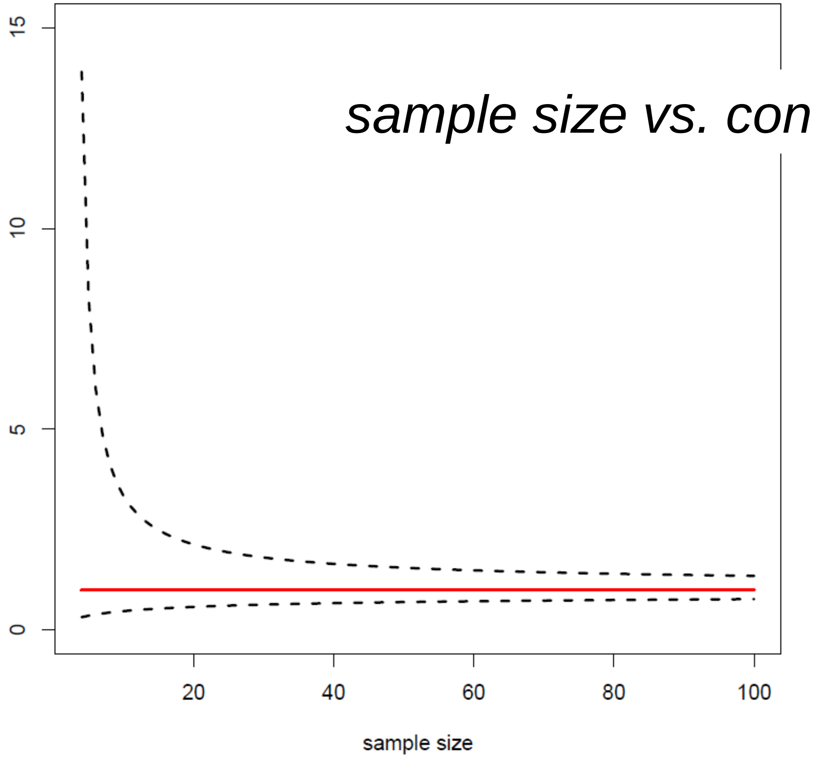
Techniques (in lecture series)

→ FWER,

→ FDR.

Issues: shrinkage

Estimation of the variance of expression levels of gene *A*, with only few samples available.

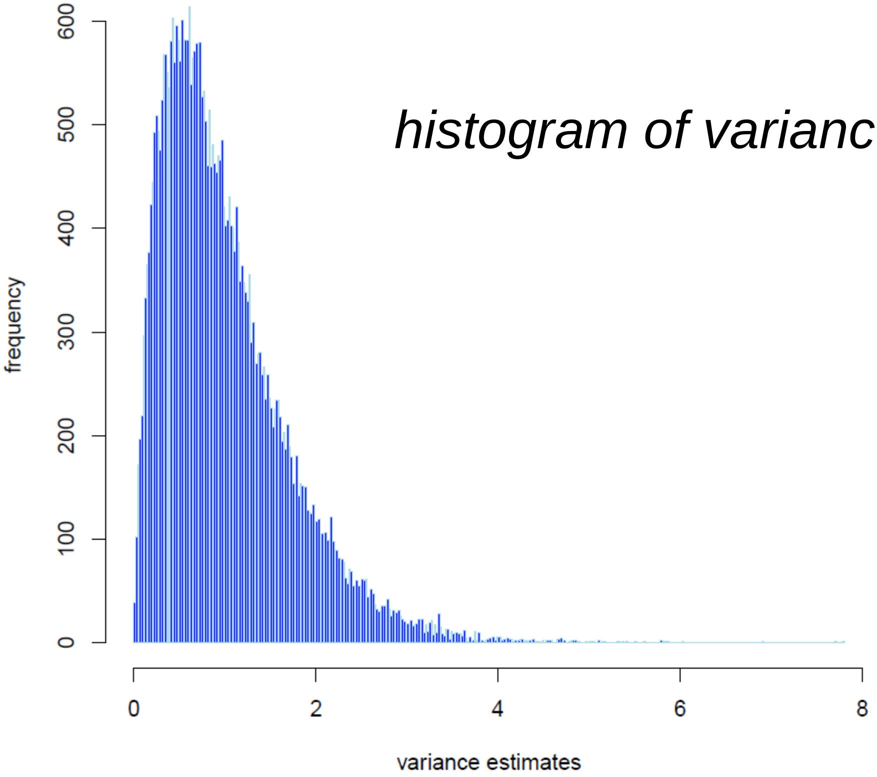


sample size vs. confidence

Few samples → large uncertainty.

Issues: shrinkage

Additional information available:
variance estimates of 40000 other genes.



histogram of variance estimates

Confidence interval of overall pooled variance estimate:
very, very small.

Issues: shrinkage

Individual variance estimate:
→ unbiased, but large uncertainly

Overall pooled variance estimate:
→ biased, but very low uncertainty

Why not exploit the strengths of boths?
E.g. by combined estimator:

$$(1-\theta)\ s^2_{\text{individual}} + \theta\ s^2_{\text{overall}}$$

The individual estimator is “shrunken” towards the overall.

Issues: shrinkage

Problem

→ low sample: highly variable estimates,
→ low-reproducibility

Shrinkage

→ traits are “comparable”,
→ borrow information across traits,
→ stabilizes estimation and improve inference.

Techniques (in lecture series)

→ Stein estimator,
→ Empirical Bayes.

Issues: penalized estimation

A common objective

predict clinical outcome from gene expression levels.

Data available:

→ a few hundred samples at best,

→ # covariates ≈ 40000.

Harrell (2001) gives the following rule-of-thumb:

> *For each continuous covariate in the model 10–20 observations are needed to detect reasonably sized effects with reasonable power.*

Where does this put us? A model with 20 genes?

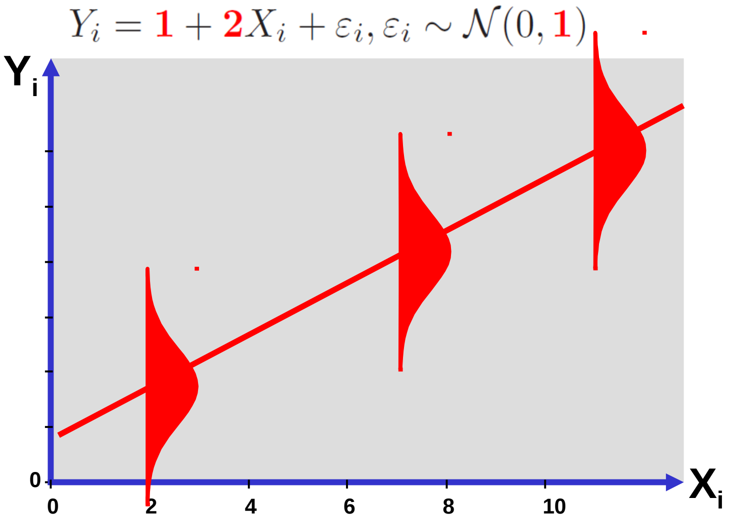
Issues: penalized estimation

Identifiability

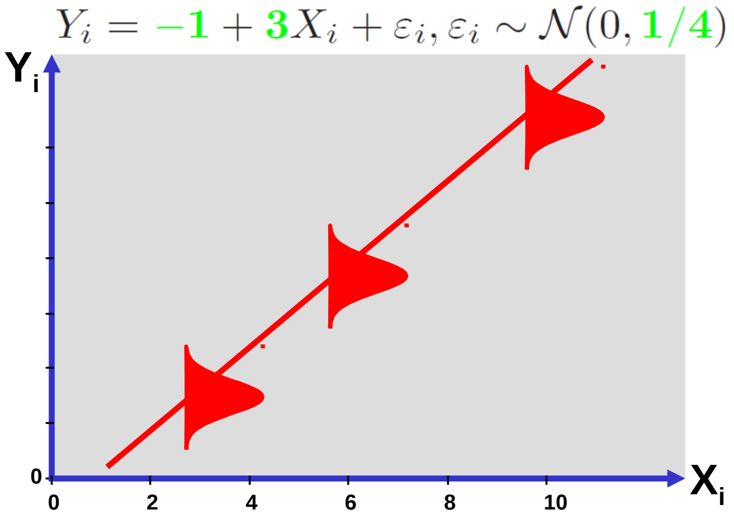
A statistical model is *identifiable* if for any two choices of the parameter θ_1 and θ_2, such that $\theta_1 \neq \theta_2$, the resulting probability distributions differ: $P_{\theta_1} \neq P_{\theta_2}$.

Fact

The linear regression is identifiable.



$Y_i = 1 + 2X_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, 1)$



$Y_i = -1 + 3X_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, 1/4)$

Issues: penalized estimation

Identifiability

Empirically, when p > n, the parameters cannot uniquely be identified from the data. That is, multiple parameter choices yield the same model.

Data available:

→ 100 samples,

→ 10000 covariates,

Then:

→ Let the first 100 covariates be linearly independent.

→ The same holds for the second 100 covariates.

→ Both sets of covariates produce a linear regression model with a perfect fit.

How do we distinguish between the two?

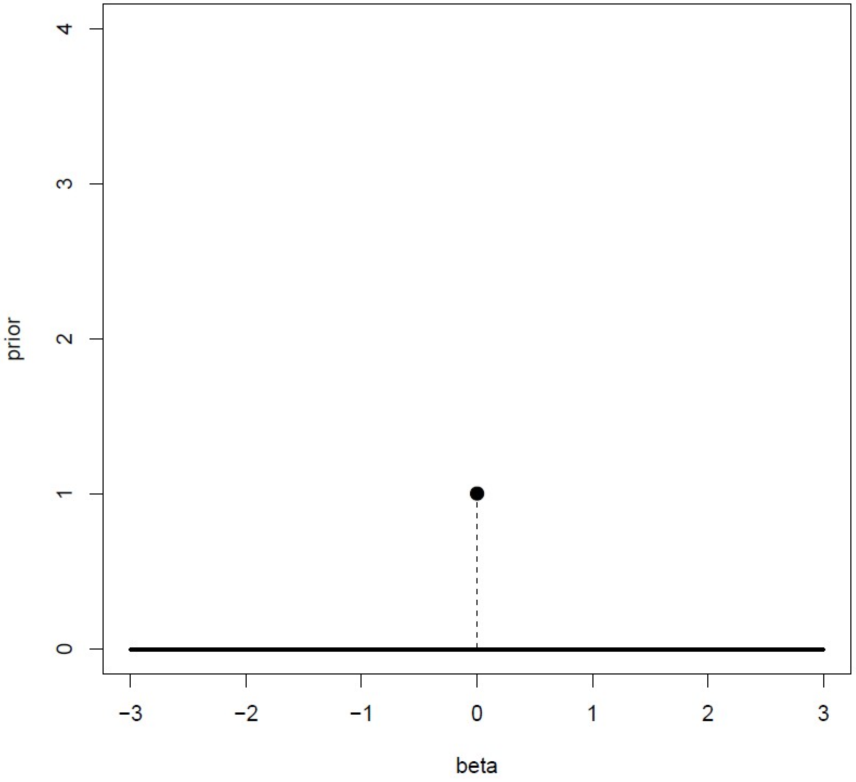
Issues: penalized estimation

Additional information may help.

In extremis
Would one know which (max) 20 covariates to include, Harrell (by his rule-of-thumb) would not object.

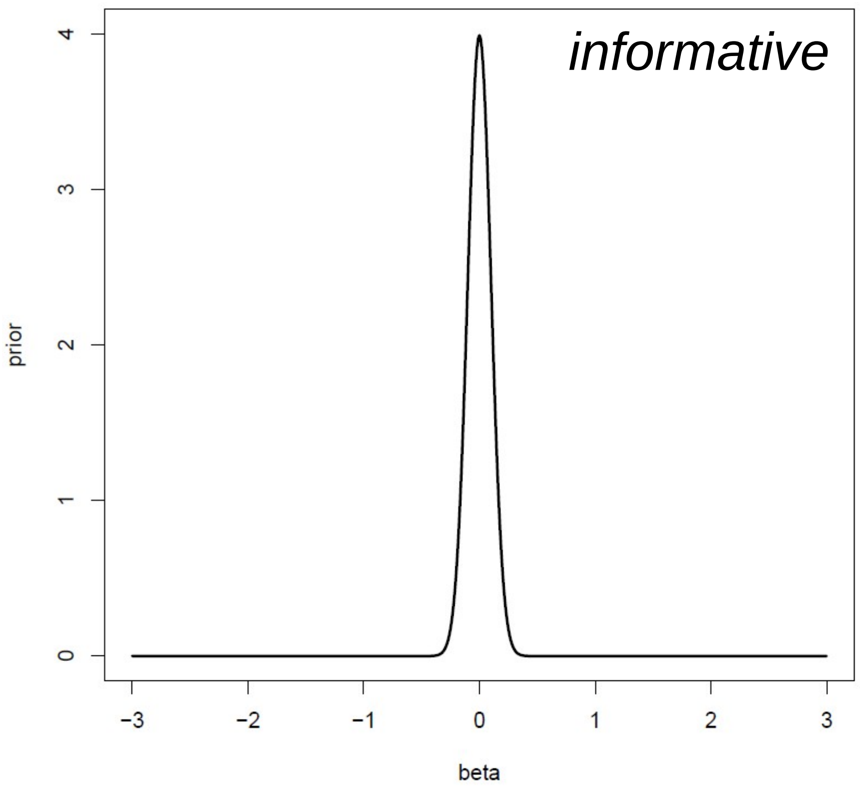
A natural way to include such information is e.g. through the specification of a prior.

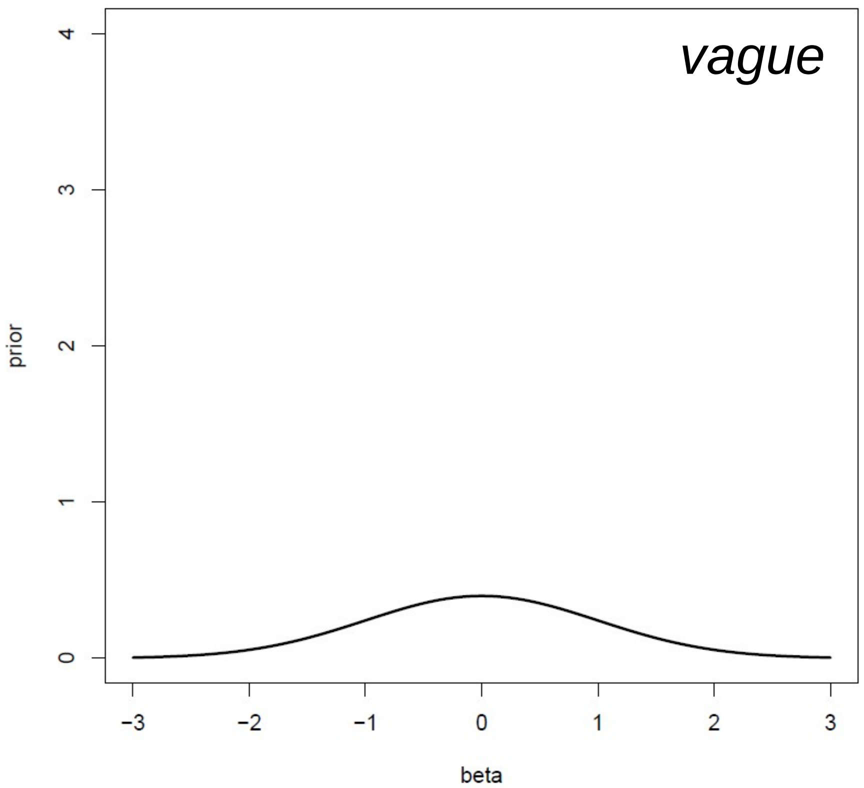
A *prior with a point mass* at zero omits the covariate from the model.



Issues: penalized estimation

Often no knowledge on relevant covariates. Data may help in the selection of the prior. E.g., very large sample size: no informative prior needed.





Issues: penalized estimation

Problem: Correlation estimates inflate

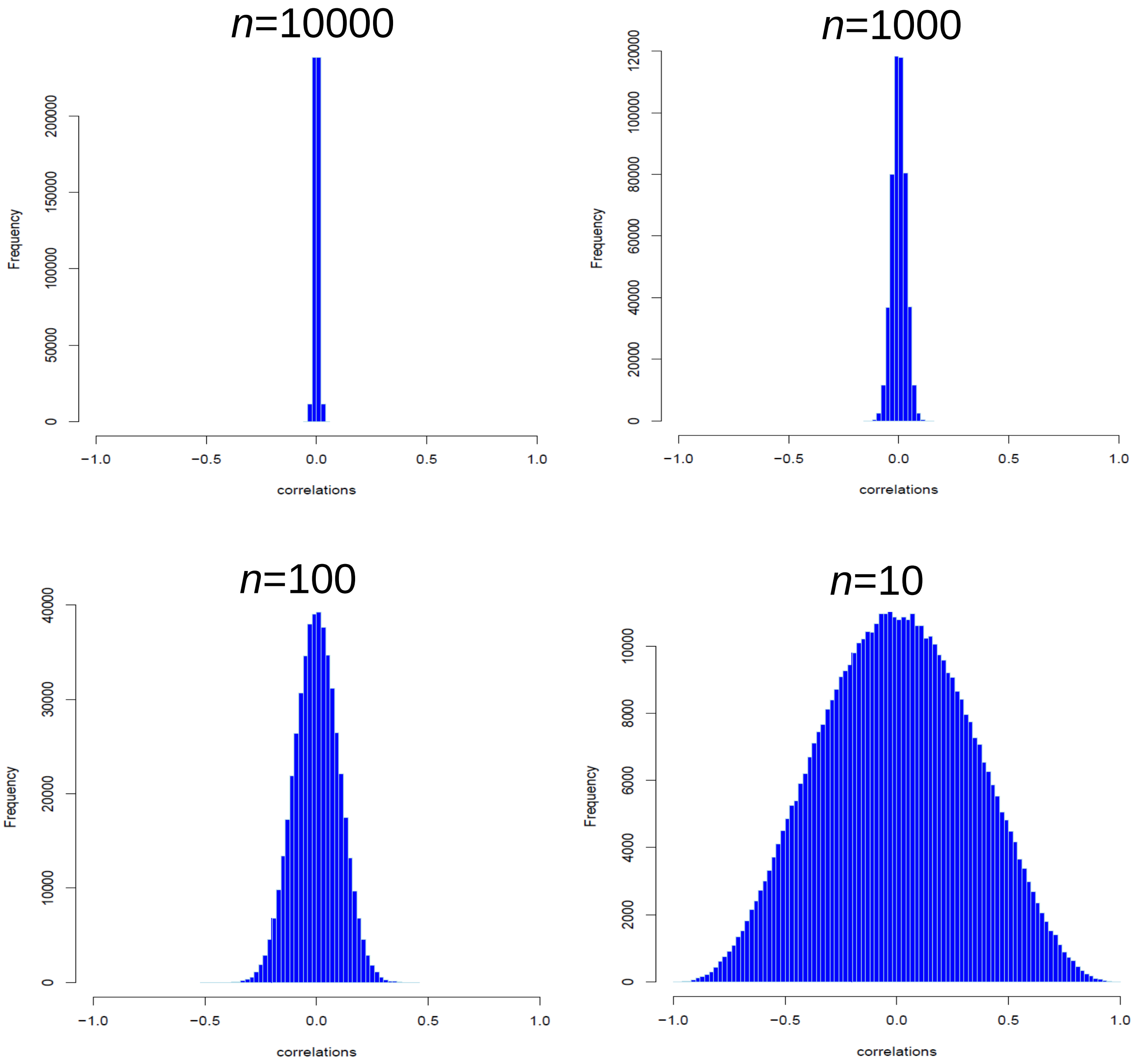
Data:

→ $Y_i \sim N(0_p, I_{pp})$

→ $p = 1000$

→ n varies

Calculate sample correlation matrix.



Issues: penalized estimation

Problem
→ few samples, many parameters,
→ estimation is frustated by p > n.

Penalized estimation
→ estimation procedures for p > n,
→ less variable but biased estimates,
→ illustrated on reconstruction of networks.

Techniques (in lecture series)
→ ridge regression,
→ lasso regression,
→ ridge and lasso (inverse) covariance estimation.

Issues: asymptotics (not treated)

Problem:

→ when *p* grows, parameters no longer fixed, e.g.:

$$\mathcal{N}_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

Various asymptotic limits discerned:

→ $n \succ p$: $n \to \infty$ and p fixed ,

→ $n \succeq p$: $n \to \infty$ and $d = \mathcal{O}(n)$,

→ $p \succeq n$: $p \to \infty$ and $n = \mathcal{o}(p)$,

→ $p \succ n$: $p \to \infty$ and n fixed .

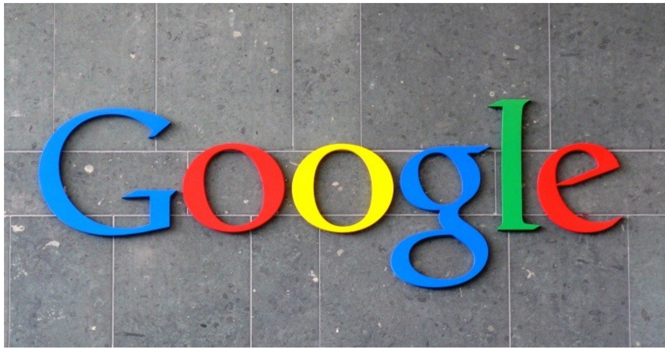
Big data vs. high-dimensional data

Big vs. high

Big data:
→ large *n*: many individuals, large sample size.
→ large *p*: information on many traits of these individuals.

Google:
→ *n* large: many people use Google software.
→ *p* large: Google registers everything these *n* do.
Similarly, Facebook, ING, et cetera.

Why:
→ many individuals available.
→ information is cheap to gain.



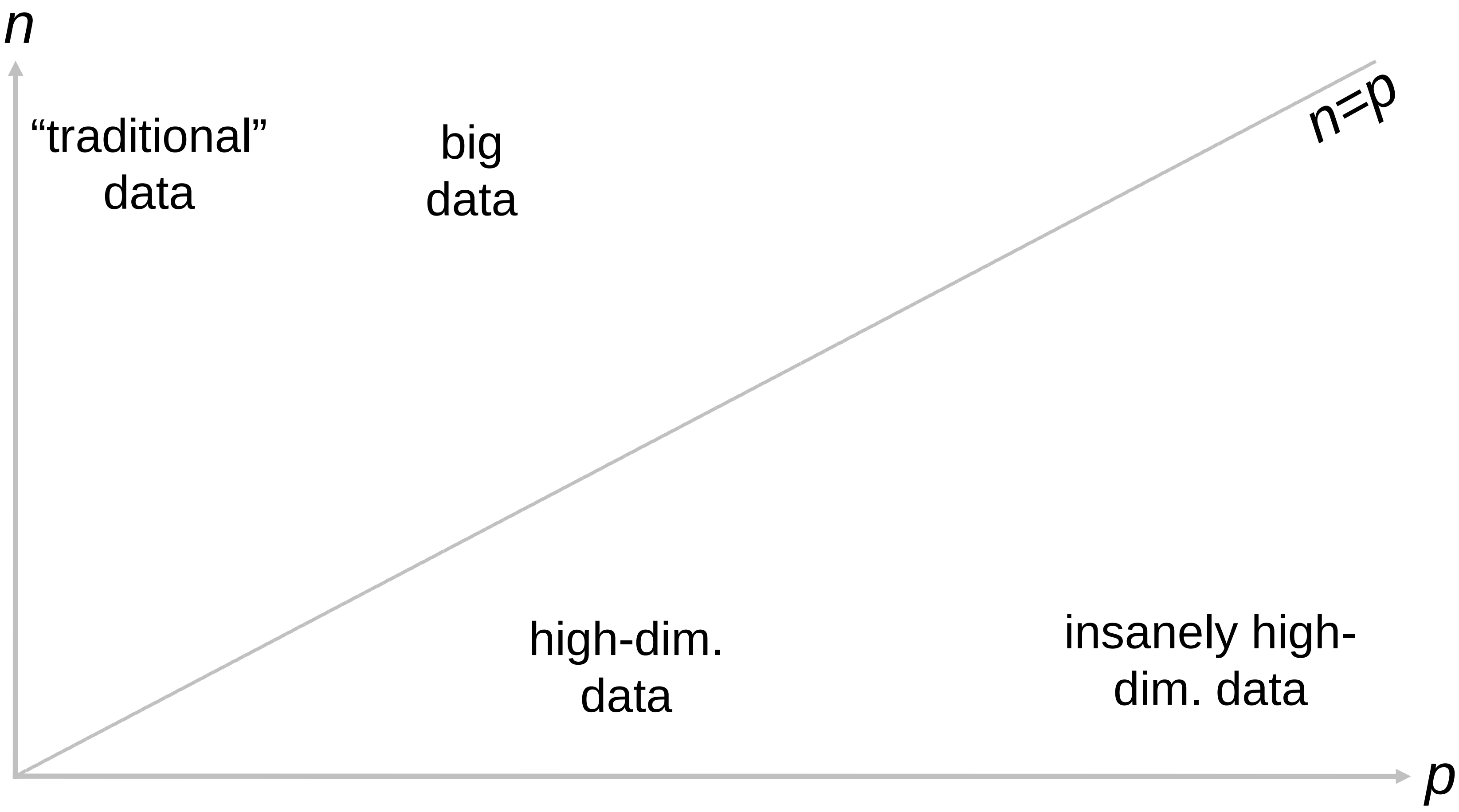
Big vs. high

High-dimensional data:
→ small *n*: few individuals, small sample size.
→ large *p*: information on many traits of these individuals.

VUmc:
→ *n* small: few cancer patients.
→ *p* large: many traits.

Why:
→ individuals with particular disease not abound.
→ information is expensive.

Big vs. high


n
“traditional” data
big data
$n=p$
high-dim. data
insanely high-dim. data
p

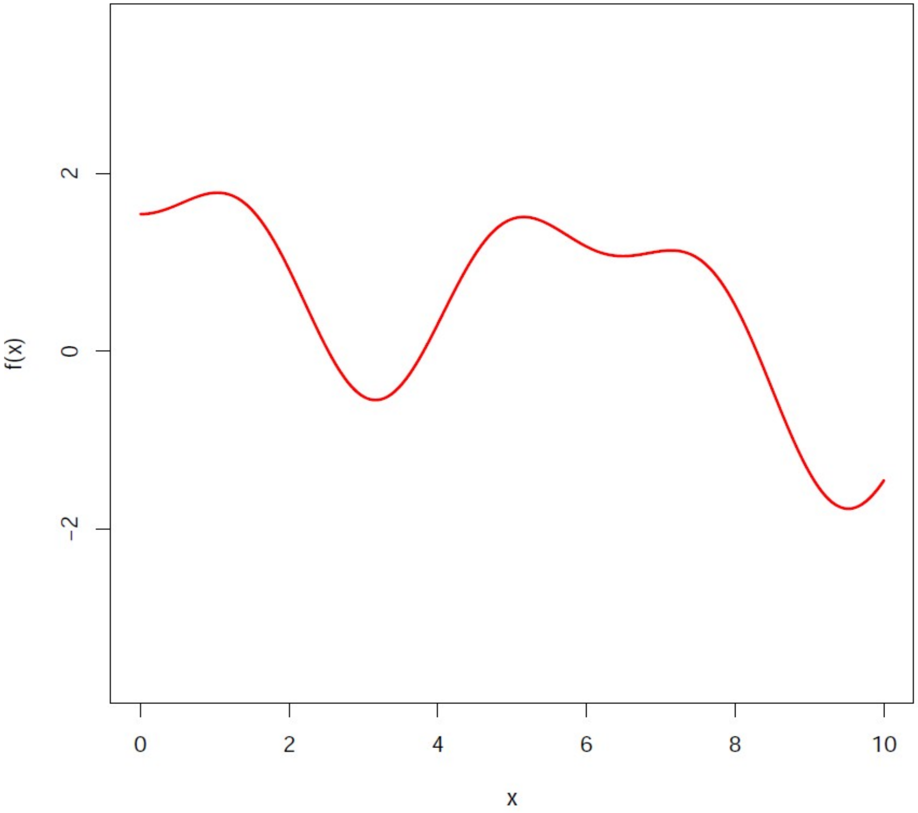
Big vs. high

Nonlinear function:
$f(x) = \sin(x) + \sin(x^2) + \cos(\exp(x^2/100))$.

Estimate f(x) from:
→ big data (n=1000),
→ high-dim. data (n=10).

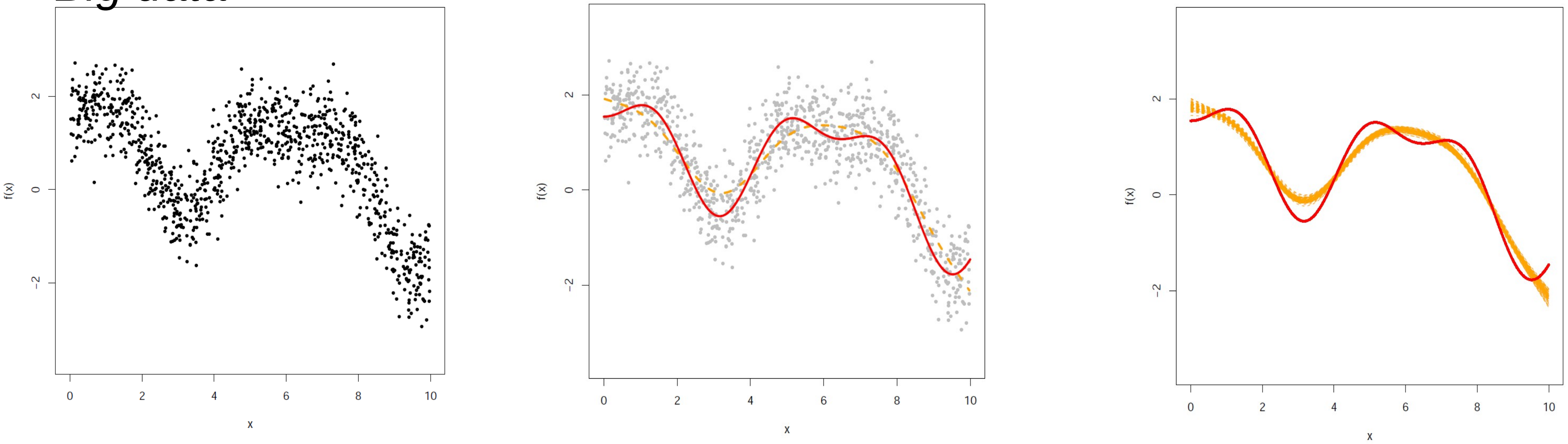
Approximate f(x) by spline with p degrees of freedom. Use $p < 7$, while f(x) requires $p >> 7$.

True function

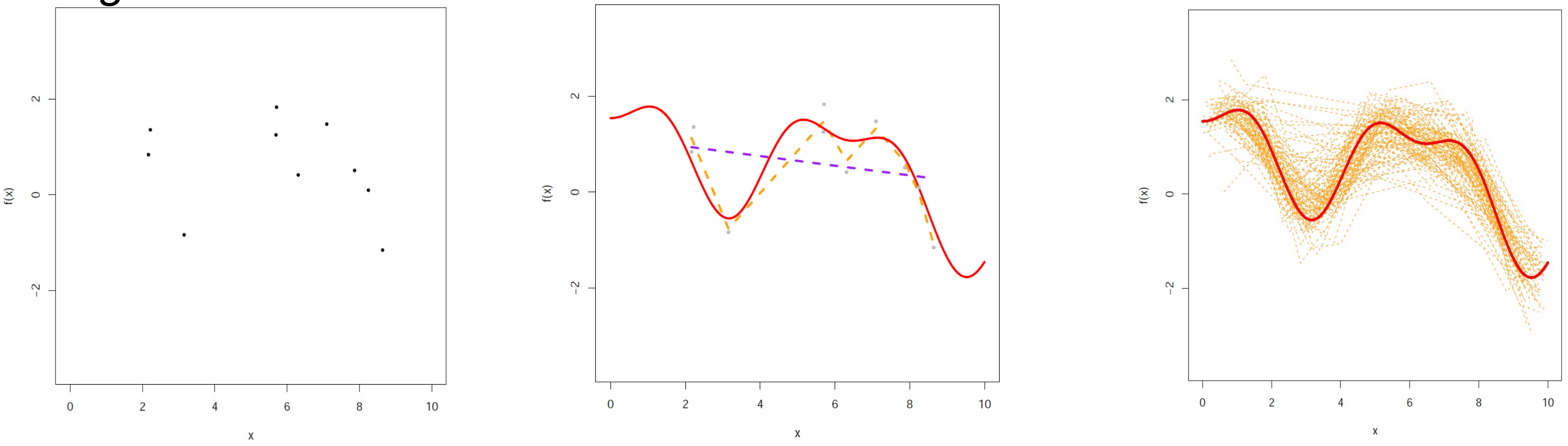


Big vs. high

Big data



High-dimensional data



Big vs. high

Big and high-dimensional data often differ in the experimental design underlying the data.

Big data

Practice
Google collects virtually anything it can gets its hands on.

Design
Observational at best

High-dim. data

Practice
VUmc financial sources limited: careful planning of experiments.

Design
Observational, but often well-controlled experiments.

Big vs. high

Big and high-dimensional data often used for different purposes.

Big data

Practice
Google optimizes advertisement revenue.

Aim
Predict behaviour of the internet user.

High-dim. data

Practice
VUmc tries to cure patients.

Aim
Understanding of the disease mechanism.

Big vs. high

Both big and high-dimensional data sometimes originate from diffuse sources.

Big data

Google measures all, but may also acquire third-party data.

High-dim. data

VUmc measures molecular and clinical traits of patients.

The data thus comes from various sources, possibly in different formats and with varying quality. To benefit from this multitude of sources is challenging.

References

References

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.D. (2003), “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”, *Bioinformatics*, **19**(2), 185-193 .

Cleveland, W.S. (1979), “Robust locally weighted regression and smoothing scatterplots”, JASA, 74(368), 829-836.

Giraud, C. (2015), *Introduction to High-Dimensional Statistics*. CRC Press.

Harrell, F.E. Jr. (2001), *Regression Modelling Strategies, with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K., Scherf, U., Speed, T.D. (2003), “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”, *Biostatistics*, **4**(2), 249-264.

Koch, I. (2013), *Analysis of Multivariate and High-dimensional Data,* Cambridge University Press.

Mayer-Schönberger, V., Cukier, K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray Publishers.

Nguyen, D.V., Bulak Arpat, A., Wang, N., Carroll, R.J. (2002), “DNA microarray experiments: biological and technological aspects”, *Biometrics*, 58, 701-717.

Quackenbush, J. (2002), “Microarray data normalization and transformation”, *Nature Genetics Supplement*, **32**, 496-501.